$3.95 / U.K. £2.75

SPECIAL ISSUE

SCIENTIFIC AMERICAN

THE COMPUTER IN THE 21ST CENTURY

10.9

Computer Networks
NICHOLAS NEGROPONTE

Computers in Education
ALAN KAY

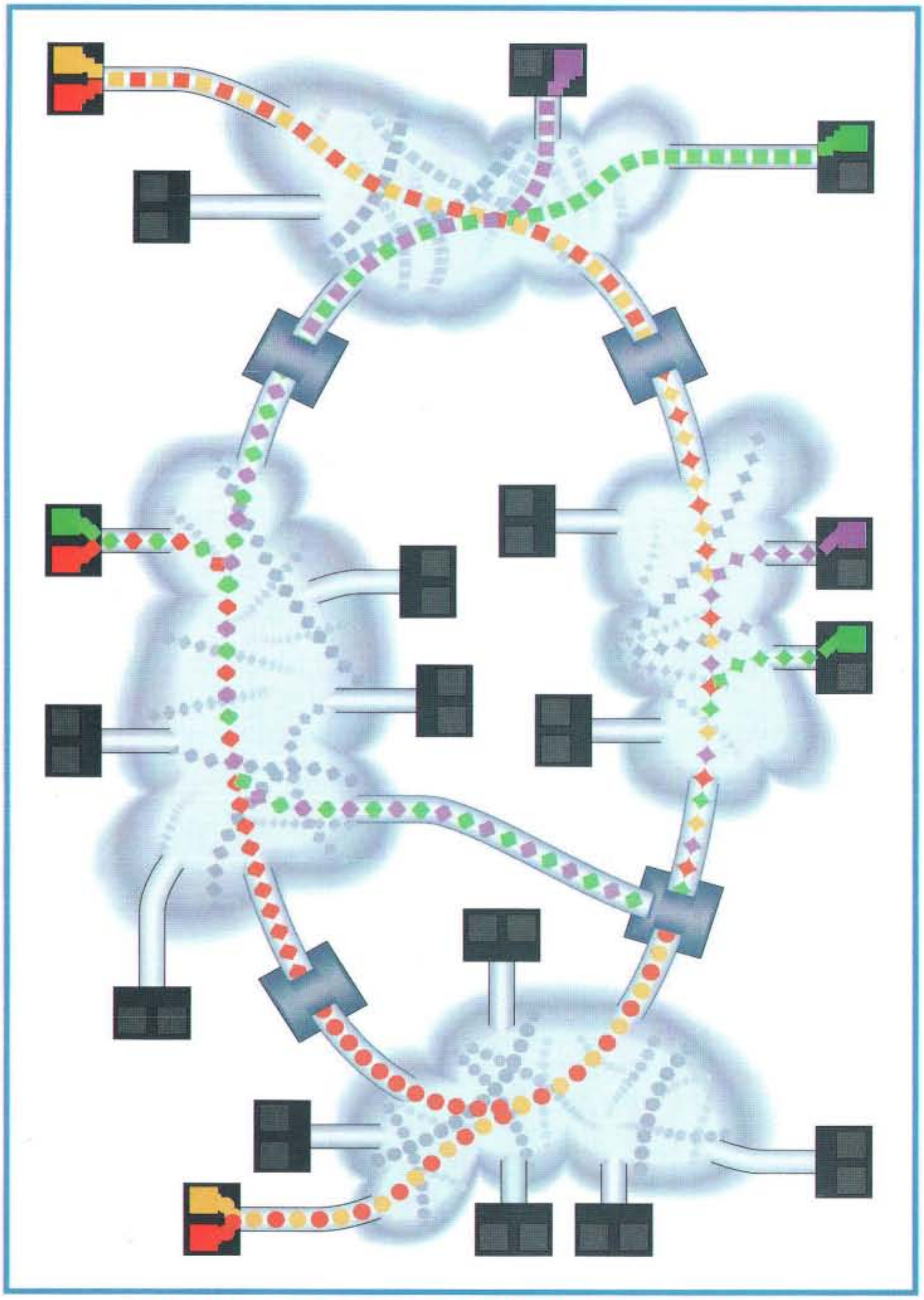
Future Computers
MARK WEISER

Networking
LAWRENCE TESLER

Civil Liberties
MITCHELL KAPOR

Government & Cyberspace
AL GORE

Computers, Networks
and the Corporation
THOMAS MALONE

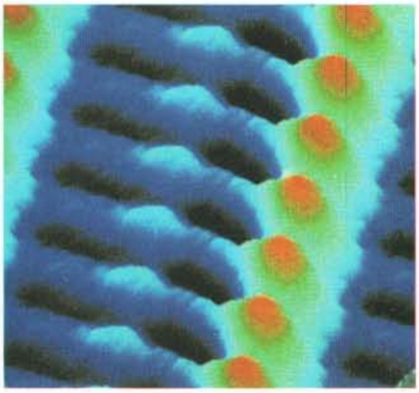


EAN
9 771048 094009 06>

The best computer articles from the pages of Scientific American magazine—updated for 1995

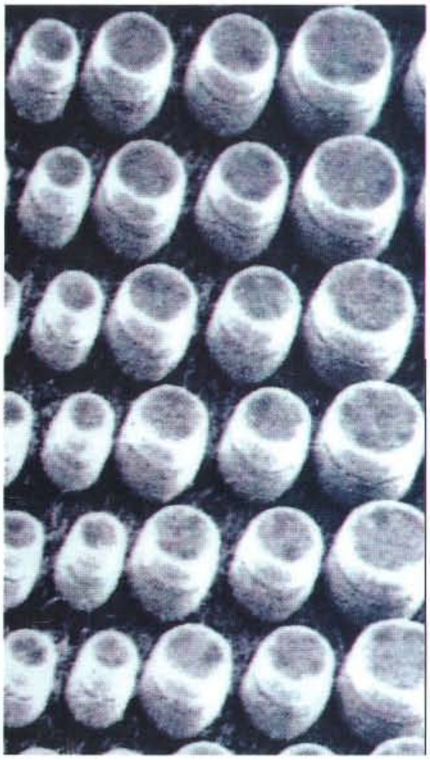
SCIENTIFIC AMERICAN THE COMPUTER IN THE 21ST CENTURY

A SPECIAL ISSUE / 1995



The Computer in the 21st Century/4
The Editors

The fusion of computer and communications technology is bringing about a major social transformation with dimensions so vast that they are difficult to discern. Scientific American *presents herewith a special issue that assesses the impact of the transformation thus far and peers ahead to glimpse what shape it may take in the 21st century.*



Networked Computing in the 1990s/10
Lawrence G. Tesler

Computers have evolved from batch processing in the 1960s to networking in the 1990s.

Communications, Computers and Networks/22
Michael L. Dertouzos

A fusion of computing and communications bids to reshape society and the economy.

Domesticating Cyberspace/30
Gary Stix

The network is the source of next-generation electronic products.

Networks/44
Vinton G. Cerf

The demands of networking will lead to new ground rules.

Lightwave Communications: The Fifth Generation/54
Emmanuel Desurvire

Erbium-doped optical fibers are revolutionizing the way signals are regenerated and transmitted.

Microlasers/62
Jack L. Jewell, James P. Harbison and Axel Scherer

Tiny lasers built into a single chip offer novel applications.

Flat-Panel Displays/70
Steven W. Depp and Webster E. Howard

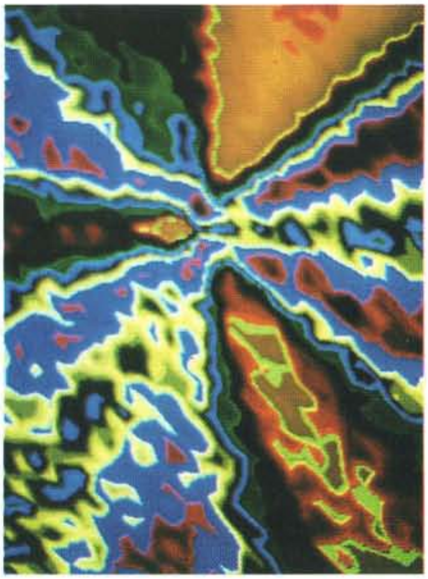
New video screens can be hung on a wall or worn on a wrist.

The Computer for the 21st Century/78
Mark Weiser

Someday computers will be ubiquitous and largely unnoticeable.

The Future of the Transistor/90
Robert W. Keyes

Millions of transistors can now be etched on a single chip. Will miniaturization continue?



Single Electronics/96
Konstantin K. Likarev and Tord Claeson

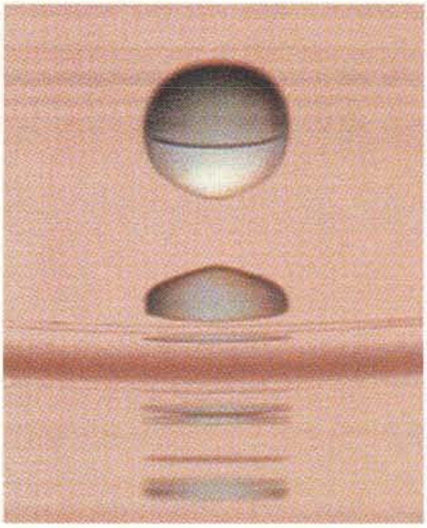
New studies of electric current may yield novel electronic devices.

Products and Services for Computer Networks/102
Nicholas P. Negroponte

Intelligent design will create networks with cunning.

When Is Seeing Believing?/110
William J. Mitchell

Digital manipulation can alter the message of a photograph.



Computers and Architecture/120
Donald P. Greenberg

How you can walk around in a structure that is yet unbuilt.

Computers, Networks and Work/128
Lee Sproull and Sara Kiesler

Computer networks will greatly affect the structure of organizations and the conduct of work.

Computers, Networks and the Corporation/140
Thomas W. Malone and John F. Rockart

Changes are coming in corporate structure and management.

Computers, Networks and Education/148
Alan C. Kay

Used wisely, the classroom computer can stimulate the mind.

Infrastructure for the Global Village/156
Al Gore

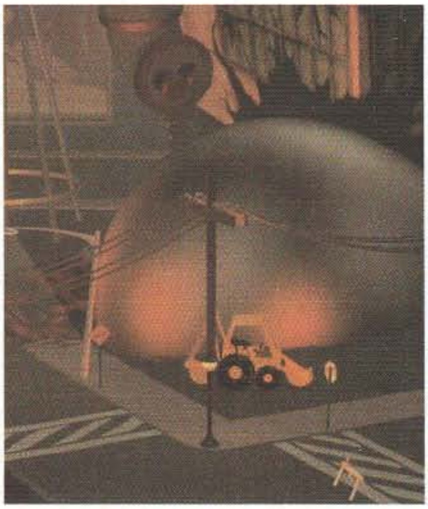
Does the information highway need government investment?

Common Law for the Electronic Frontier/160
Anne W. Branscomb

Cyberspace is an electronic frontier where the law is in doubt.

Quantum Cryptography/164
Charles H. Bennett, Gilles Brassard and Artur K. Ekert

Quantum mechanics may make it possible to exchange messages in absolute secrecy.



Civil Liberties in Cyberspace/174
Mitchell Kapor

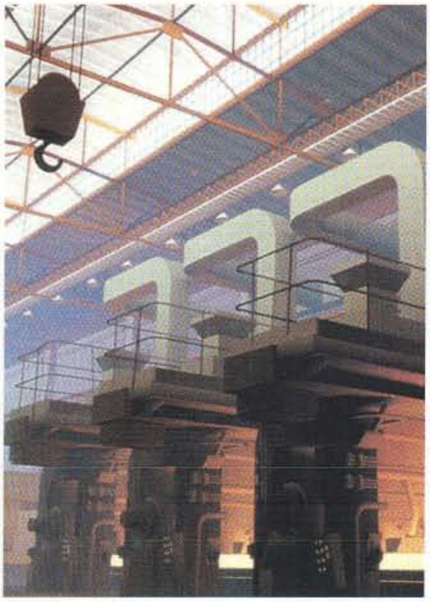
The line between computer hacking and crime is blurry.

The Risks of Software/180
Bev Littlewood and Lorenzo Strigini

Uncertainty in the reliability of software must be taken into account in critical systems.

Wire Pirates/186
Paul Wallich

The Internet depends on mutual trust and forbearance, traits that appear to be increasingly falling victim to villains.



SCIENCE AND BUSINESS

When Cells Divide/76
Making space for the next wave of wireless communications.

Heads in the Cloud/126
Can programmers entice consumers into the net?

Dr. Big Brother/172
Will health care reform make everyone's life an open book?

Tap Dance/179
Keeping communications networks safe for bugging.

Surreal Science/195
Virtual reality finds a place in the classroom.

SCIENCE AND THE CITIZEN

Make, Model and.../69
A privacy advocate puts license plates on-line.

Electronic Communication/196
Is it history or just E-mail?

ANALYTICAL ECONOMIST

A Digital Fix for the Third World?/43
Technology in developing nations.

LETTER FROM THE PUBLISHER

You're reading one of the first magazines produced without the usual film prepared in advance for offset printing, and it's still only the 20th century! It seems fitting that a magazine about the most important new computing would itself be generated by new computer-based technology.

Woodblocks, in fifth-century China, marked the world's first mass-production printing—the first on-ramp to the Information Superhighway. Photographic film began to be used to create the plates required for printing in 1847, two years after the launch of *Scientific American.* Now—here—you have a magazine whose images have been transmitted in the form of digital data direct from the computers of the magazine's editors, art and production staff to the pressroom itself without any film intermediary.

All 200-odd pages of *The Computer in the 21st Century* were published by the new computer-to-plate (CTP) technology, as it is termed, and printed by the web-offset printing process. Significantly, the inks are soy-based and environment-friendly (and used in all copies of *Scientific American*), and the paper is chlorine-free. Also, most of the issue is being published electronically in *Scientific American's* America Online "newsstand" (keyword: SciAm). Comments? Post them in the "Reader to Reader" on-line message board so that all can participate; either *Scientific American's* Vice President/Production, Richard Sasso, or I will read and respond.

To publish this kind of new material, which, in quality and quantity, would constitute a fair-sized, costly book—and to do so for a reader price of just a few dollars—would be impossible without the commercial support of the sponsor, Apple Computer, and Apple's advertising agency, BBDO/Los Angeles. Apple's advertising pages were supplied as digital data for CTP printing. The medium is indeed the message.

Please don't let all this innovation detract from the experience of reading and appreciating the articles, columns and information that follow. Antique though it may be, ink on paper still stimulates and inspires!

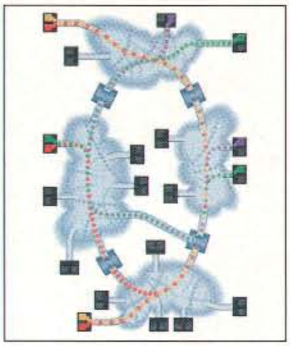
John Moeling
President and Publisher
Scientific American, Inc.

NOTE ON THE PRODUCTION OF THIS ISSUE

Using computer-to-plate-technology, *Scientific American* supplied editorial text and advertising materials to our printer, the Mattoon, Illinois, division of R. R. Donnelley & Sons Company, as complete digital data to image plates electronically for web-offset printing, without any intermediate film stages. No film was used to prepare and produce any of the 200 editorial and advertising pages.

Since October 1994, *Scientific American* has been committed to the regular use of CTP technology. For the production of this special anthology, we supplied complete editorial and advertising pages, produced using the Apple Macintosh, with editorial illustrations and text made up in Adobe Illustrator and Quark XPress. Along with the low-resolution proxy files of the illustrations, we sent the high-resolution color separations as produced by the magazine's prepress vendors (Capitol Engraving, Nashville, Tennessee, and Black Dot Graphics, Crystal Lake, Illinois) on their proprietary systems. The size of the total editorial and advertising file was three gigabytes. All 21 Apple Computer supplied advertising pages were delivered through its agency BBDO/Los Angeles and its vendor (Effective Graphics, Inc., Compton, California) as digital data for PostScript imaging. No conversions were made. This would not have been possible without the complete cooperation of the agency's media and production departments, the prepress vendors and our printer. 3M Rainbow and Kodak Approval digital proofs were used in the process. Plate imaging was accomplished by sending PostScript data to a Creo 3244 Platesetter (from Creo Products, Inc., Burnaby, B.C.), with which Hoechst Celanese N90 plates were imaged. The magazine was printed by web offset to a press count of 300,000 copies on 40-lb Bowater Jet Blade chlorine-free paper and 100-lb Champion Javelin paper, using NaturaLith, a vegetable-based printing ink manufactured by SunChemical. —*Richard Sasso*
(E-mail: richsasso@sciam.com)

SCIENTIFIC
AMERICAN®



THE COVER sounds the theme of this issue with a representation by Michael Goodman of the Internet—a network of networks.

THE COMPUTER IN THE 21ST CENTURY

SCIENTIFIC AMERICAN *The Computer in the 21st Century* is published by the staff of SCIENTIFIC AMERICAN, with project management by:

EDITOR IN CHIEF: John Rennie

MANAGING EDITOR: Michelle Press

PROJECT EDITOR: James T. Rogers

PROJECT ART DIRECTOR: David Starwood

COPY: Maria-Christina Keller, *Copy Chief;* John P. Timpane

PRODUCTION: Richard Sasso, *Associate Publisher/Vice President, Production;* William Sherman, *Production Manager; Managers:* Carol Albert, *Print Production;* Janet Cermak, *Makeup & Quality Control;* Tanya DeSilva, *Prepress;* Silvia Di Placido, *Special Projects;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Ad Traffic: Carl Cherebin; Carey S. Ballard; Kelly Ann Mercado

CIRCULATION: Lorraine Leib Terlecki, *Associate Publisher/Circulation Director;* Katherine Robold, *Circulation Manager;* Joanne Guralnick, *Circulation Promotion Manager;* Rosa Davis, *Fulfillment Manager*

ADVERTISING: Kate Dobson, *Associate Publisher/Advertising Director.* OFFICES: NEW YORK: Meryle Lowenthal, *New York Advertising Manager;* Randy James, Rick Johnson, Elizabeth Ryan, Timothy Whiting. CHICAGO: 333 N. Michigan Ave., Suite 912, Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager.* WEST COAST: 1554 S. Sepulveda Blvd., Suite 212, Los Angeles, CA 90025; Lisa K. Carden, *Advertising Manager;* Tonia Wendt. 235 Montgomery St., Suite 724, San Francisco, CA 94104; Debra Silver. CANADA: Fenn Company, Inc. DALLAS: Griffith Group

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie M. Beaumonte, *General Manager;* Constance Holmes, *Manager, Advertising Accounting and Coordination*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017-1111
(212) 754-0550

CHAIRMAN AND CHIEF EXECUTIVE OFFICER: John J. Hanley

CO-CHAIRMAN: Dr. Pierre Gerckens

CORPORATE OFFICERS: John J. Moeling, Jr., *President;* R. Vincent Barger, *Chief Financial Officer;* Robert L. Biewen, *Vice President*

DIRECTOR, ELECTRONIC PUBLISHING: Martin Paul

PRINTED IN U.S.A.

THE ILLUSTRATIONS

Page	Source
4	Chas. Addams, © 1952, 1980, The New Yorker Magazine, Inc.
5	Imperial College, London
7	Comstock, Inc.
8	Magellan Systems
9	Steven E. Sutton/Duomo
11	IBM (*top left*), Digital Equipment Corporation (*top right*), Ed Keating (*bottom left*), Apple Computer Inc. (*bottom right*)
12	IBM (*top photo insets*), Cray Research, Inc. (*bottom photo inset*); Cray photo by Paul Shambroom
17	IBM Corporation (*left*), Fujitsu (*right*)
18	IBM Corporation
20	Ed Keating
21	© 1970 Universal Pictures, Inc.
22–23	John Telford, University of Utah
24	Stanley Rowin
25	Nynex; photo by Laurie Belisle; image by Hubert C. Delaney, © 1986
26	Kent Kallberg
27	IBM Corporation
28	Joe Ruskin
29	Lillian Schwartz and Rick Becker
30–31	John McGrail
32–33	George Retseck
35	Merry Alpern
37	Geoffrey Wheeler/Black Star (*left*), Microsoft Corporation (*right*)
44–45	© 1991 John McGrail
46	Michael Goodman
47	Johnny Johnson (*top*), © 1991 John McGrail (*bottom*)
48–51	Michael Goodman
52–53	Brian K. Reid, Digital Equipment Corporation
54–55	Emmanuel Desurvire
56–59	Ian Worpole
60	Neal S. Bergano, AT&T Bell Laboratories
61	Terry W. Cline, AT&T Bell Laboratories
63	Axel Scherer
64, 66–67	Michael Goodman
68	Axel Scherer
69	Jessica Boyatt
70–71	John S. Foster, Jane Frommer and Jacquelin K. Spong, IBM Thomas J. Watson Research Center (*left*), Michael Goodman (*right*)
72	Photonics (*top left*), Planar Systems (*top right*), Michael Goodman (*bottom*)
73	Michael Goodman
74	Laurie Grace
75	Carolco
77	U.S. Department of Commerce
79	Matthew Mulbry
80	Ian Worpole
85	Ian Worpole; Matthew Mulbry (*photo inset*)
86–89	Matthew Mulbry
90–91	IBM Corporation
94–95	IBM Corporation
97	Lester Lefkowitz
98	Ian Worpole
99	Ian Worpole (*left, a–d*), Lester Lefkowitz (*right, a–d*)
100–101	Ian Worpole
103	Quasada/Burke; image courtesy of Media Laboratory, Massachusetts Institute of Technology
105	Quesada/Burke; image by Thomas Van Sant, Inc.
108–109	Media Lab, M.I.T.
110–111	Original photograph by AP/World Wide Photos (*left*), image from Paul Higdon/*New York Times* (*center*), image from Angela Perkins (*right*), color manipulations by Jack Harris/Visual Logic (*top center, top right* and *bottom*)
112	Reuters/Bettmann (*top*), Department of Defense Photo (*bottom*)
117	Original photography by NASA; digitally manipulated version courtesy of Time, Inc., Picture Collection
118	Lincoln photograph by Alexander Gardner, Bettmann Archive; Monroe photograph courtesy of Personality Photos, Inc.; digital images by Jack Harris/Visual Logic
119	Image by Wade Hokoda
120–121	Charles Gwathmey (*left*), Donald P. Greenberg (*center top*), Keith Howie, Ben Trumbore, Donald P. Greenberg (*center bottom*), Gwathmay Siegel & Associates Architects (*right*)
122	Donald P. Greenberg
123	K. Howie and B. Trumbore (*top right*), John Wallace, Eric Chen and Michael Cohen (*bottom right*)
124–125	K. Howie, Paul Boudreau, Eric Haines and John Wallace (*top left, bottom*), Gabor Kiss (*top center*), E. Chen, Stuart I. Feldman and Julie O'Brien (*top right*)
126	General Magic
127	Motorola, Inc.
129	Jason Goltz
130–134	Laurie Grace
139	Marc Skinner; image by Nathaniel Borenstein, Bellcore, Inc.
141	Paolo Koch, Photo Researchers, Inc. (*top*), Brett Froomer, The Image Bank (*bottom*)
142	Rosenbluth Travel, Inc.
143	Frito-Lay, Inc.
144	Kurt Salmon Associates (*top*), Edward Bell (*bottom*)
146	JoMarie Fecci
147	Comstock, Inc./Russ Kinne
148–149	Warren Faubel/Black Star
151–152	Warren Faubel/Black Star
153	National Geographic Society
154	Walter Bender, © 1984 M.I.T.
155	Warren Faubel/Black Star
157–158	White House Photo Office
165	Robert Prochnow
166	Courtesy of David Kahn, © 1983 Macmillan Publishing Company
167–171	Michael Goodman
173	Wally Man Namee/Sygma
175	David J. Sams
177	CompuServe
179	Mykotronx, Inc.
180–181	D. Aubert/Sygma
185	Nuclear Electric
186–187	Patricia J. Wynne
188	Jason Goltz
189	Jared Schneidman Design
190	Stephanie Rausser
191	Jared Schneidman Design
192	Chris Usher/Black Star
193	Jared Schneidman Design
194	Stephanie Rausser
195	R. Bowen Loftin
196	*The Iran-Contra Scandal: A Declassified History,* edited by Peter Kornbluh and Malcom Byrne, © 1993 The New Press

D

Scientific American The Computer in the 21st Century (ISSN 1048-0943), Special Issue Volume 6, Number 1, 1995, published by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017-1111. Copyright © 1995 by Scientific American, Inc. All rights reserved. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. To purchase additional quantities: 1 to 9 copies: U.S. $3.95 each plus $1.00 per copy for postage and handling (outside U.S. $2.00 P & H); 10 to 49 copies: $3.95 each, postpaid; 50 copies or more: 20 percent discount. Send payment to Scientific American, Dept. CO, 415 Madison Avenue, New York, N.Y. 10017-1111. Canadian GST No. R127387652.

The Computer in the 21st Century

In 1991 the fusion of computing and communications was still a dream. Today it is a powerful social, economic and cultural reality

by the Editors

A cartoon published some years ago showed two paleontologists standing beside their Jeep, which is parked in a vast, clawed footprint. It is evident from the caption that they have no idea of what is surrounding them. The image stands as a good model for the state of mind of people caught in the middle of a major social transformation. Its dimensions are so encompassing that they are extremely hard to discern. Imagine the citizens of the 18th century trying to envision the shape of a future that would include electrical power, telecommunications, jet transportation and biotechnology. We who are alive at the end of the 20th century are having the same trouble discerning the impact of an evolutionary force that is reshaping our world: the fusion of computer and communications technology.

To help managers, entrepreneurs, administrators—and everyone else—see that process more clearly, *Scientific American* in 1991 published a single-topic issue, "Communications, Computers and Networks." Since then, the forces of innovation that are transforming our society, our economy and our culture have swept beyond the laboratory and the specialized communities that the technologies originally served and are propelling us fast-forward into the future. So the Editors have decided it is time to update the issue. It appears here as "The Computer in the 21st Century."

This special edition consists of the original articles, some revised by their authors. Accompanying them are the most significant articles concerning these events we have published during the intervening months, as well as news stories that also mark the forward movement of the trend from our "Science and the Citizen" and "Science and Business" columns.

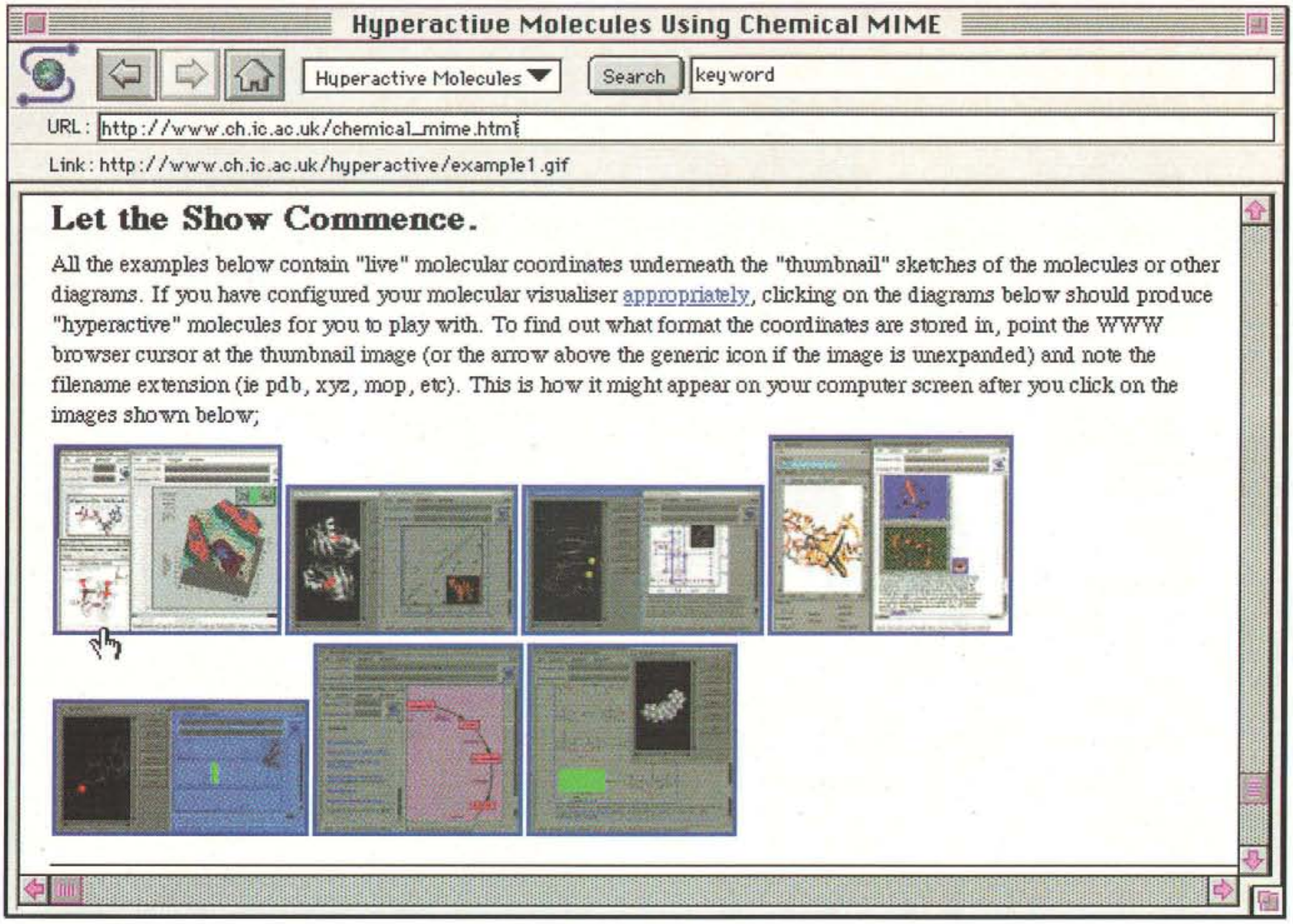
For all of its photonic speed, this revolution has been a long time in coming. Its first prophets were not McLuhan, Minsky or Gerbner, but Nathaniel Hawthorne and Karl Marx. "Is it a fact...," wrote Hawthorne, in *The House of Seven Gables,* "that by means of electricity, the world of matter has become a great nerve, vibrating thousands of miles in a breathless point of time. Rather, the round globe is a vast head, a brain, instinct with intelligence!"

Marx also speculated about the impact that global communication would have on the industrial economies. (Now that communism is safely dead, one can afford to note that Marx did make a number of acute observations about capitalism, particularly about the relation between economic evolution, science and technology.) In *Das Kapital,* the sage of the British Museum declares that the telegraph, steamships and railroads would accelerate the circulation of capital, thereby promoting the rapid growth of the industrial economies.

In this century the revolution's first prophet was Aldous Huxley. In *Those Barren Leaves,* a character observes: "Cheap printing, wireless telephones, trains, motor cars, gramophones and all the rest are making it possible to consolidate tribes, not of a few thousands, but of millions." But it was Marshall McLuhan, the University of Toron-

Drawing by Chas. Addams; © 1952, 1980; The New Yorker Magazine, Inc.

"Well, I don't see any point in looking any farther. It was probably just one of those wild rumors."



URL: http://www.ch.ic.ac.uk/chemical_mime.html

Link: http://www.ch.ic.ac.uk/hyperactive/example1.gif

Let the Show Commence.

All the examples below contain "live" molecular coordinates underneath the "thumbnail" sketches of the molecules or other diagrams. If you have configured your molecular visualiser appropriately, clicking on the diagrams below should produce "hyperactive" molecules for you to play with. To find out what format the coordinates are stored in, point the WWW browser cursor at the thumbnail image (or the arrow above the generic icon if the image is unexpanded) and note the filename extension (ie pdb, xyz, mop, etc). This is how it might appear on your computer screen after you click on the images shown below;

ACTION ON THE INTERNET is exemplified by this view of a user's computer screen. The Internet, started by the Defense Department as ARPAnet, now encompasses a large number of computer networks and some 25 million users.

to sociologist, who resynthesized the idea and presented it at the right moment. With such phrases as "the global village" and "the medium is the message," McLuhan focused the attention of the Age of Aquarius on the fact that television, radio, the telephone and the jet aircraft had woven the postwar world into a single community. (The keepers of McLuhan's intellectual heritage say the Canadian sociologist coined the term "global village" while he was contemplating Hawthorne's passage.)

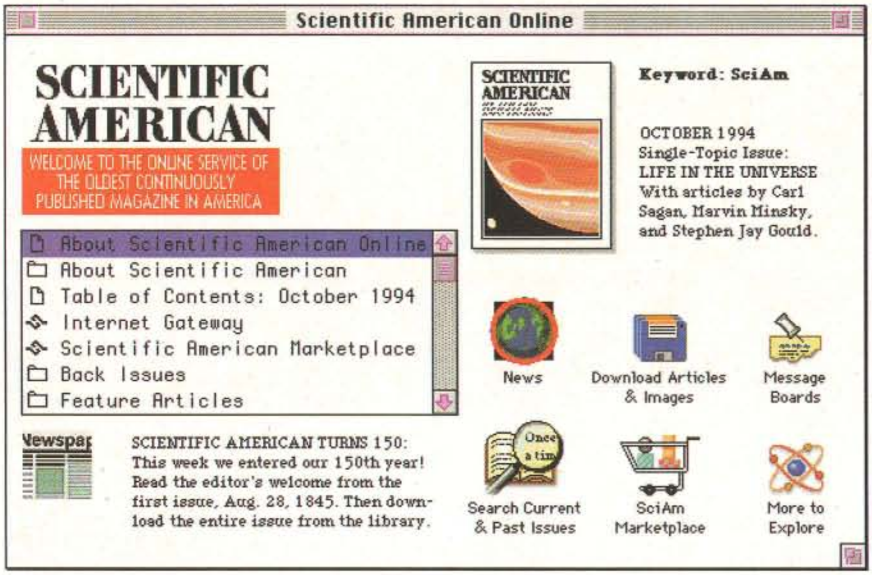
The lack of a clear economic path forward also marks this revolution as unusual. In the past few months, several mergers and alliances formed between corporations in broadcasting, cable television and telephonic communications (both wired and cellular) have collapsed or at least pulled off the road for drastic repairs. QVC's attempt to merge with CBS fell apart when another company bid for QVC. Bell Atlantic and TeleCommunications abandoned their merger when a decision by the Federal Communications Commission about cable rates chilled projections of future revenues (there was a flurry of reportage about corporate culture clashes as well).

Yet the transformation of economy, society and culture by the fusion of the technologies of computing and communications seems as inevitable as water running downhill. Four tributaries feed the flood. The computer has evolved from a roomful of electronics tended by acolytes in white coats, capable of executing perhaps 500,000 instructions per second, into a chip one square centimeter in size that works 20 times faster. Such a minute bit of petrified intelligence can even be embedded in a device such as an automobile engine or a VCR. Copper wire carrying 64,000 to up to a million bits per second is now being replaced by optical fiber with a capacity of millions and eventually billions of bits per second. Compression

Ubiquitous Computing

Computers, according to Mark Weiser [see "The Computer for the 21st Century," page 78], will someday be as ubiquitous and as woven into the fabric of daily life as street signs are now. Updating his article, Weiser writes: "Since 1991, we have continued to invent ubiquitous computing. This is long-term work, perhaps decades to completion, yet there is commercial success along the way. For instance, Xerox has sold with great success the Liveboards described in my 1991 article.

"An important development not foreseen in 1991 is the growing use of computers for casual human interaction. Most network traffic today is people talking to one another. Ubiquitous computing work has incorporated this trend. For instance, our Jupiter project links virtual and real rooms in a user-constructed, on-line environment."


Scientific American Online
About Scientific American Online
About Scientific American
Table of Contents: October 1994
Internet Gateway
Scientific American Marketplace
Back Issues
Feature Articles
Newspap
News
Download Articles & Images
Message Boards
Search Current & Past Issues
SciAm Marketplace
More to Explore

SCIENTIFIC AMERICAN
WELCOME TO THE ONLINE SERVICE OF THE OLDEST CONTINUOUSLY PUBLISHED MAGAZINE IN AMERICA

Keyword: SciAm

OCTOBER 1994
Single-Topic Issue:
LIFE IN THE UNIVERSE
With articles by Carl Sagan, Marvin Minsky, and Stephen Jay Gould.

SCIENTIFIC AMERICAN TURNS 150:
This week we entered our 150th year! Read the editor's welcome from the first issue, Aug. 28, 1845. Then download the entire issue from the library.

ON-LINE VERSION of *Scientific American* appears on the screen of a subscriber to America Online, one of several commercial bulletin boards. Serving some five million subscribers, the bulletin boards offer computer users access to many databases, periodicals and shopping services.

technology can multiply such numbers by a factor of two or three [see "Lightwave Communications: The Fifth Generation," by Emmanuel Desurvire, page 54]. The flat-panel display has liberated the computer from the desktop [see "Flat-Panel Displays," by Steven W. Depp and Webster E. Howard, page 70]. And relatively intuitive instructions, icons, windows, desktop metaphors and the mouse have replaced strings of alphanumeric instructions that only a programmer could love.

The new technology has spawned two trendy offspring: multimedia and virtual reality. Multimedia is the first attempt to orchestrate sound, visual images (usually moving) and text into a flow of knowledge or entertainment whose course the user can affect. Eventually, such packages of knowledge and entertainment will move over the network. Virtual reality simulates physical experience through stereographic visual displays, tactile sensation and response to gestures. The induced experience can be vivid, but will the technology ever be more than an excellent—and expensive—way to train a pilot, a professional driver or a surgeon? Will it, or multimedia, become a new channel of artistic expression—or an ultimate form of pornography? (Pornographic multimedia CD-ROMs are already reportedly selling tens of thousands of copies.)

Humble as it is, a computer equipped with a modem and communications software provides a better sense of the look and feel of the future. Connected to the telephone system, it abolishes space, making everywhere here. You can talk to people all over the world or browse in a geographically distant library or database. This space creates its own time, whose pace lacks both the annoying nowness of a ringing telephone or the frustrating slowness of snail mail. Most important, you can get knowledge, information or entertainment when you want it, not when someone else decides to provide it.

Millions of people already live in this future. Financiers trade in a market that never sleeps. *The Economist* has estimated that the turnover through digital channels of foreign exchange in a single week equals the U.S. national debt of some $4 trillion. Executives in multinational corporations make decisions based on management information systems fed by subsidiaries all over the planet.

Members of the world's scientific community live in a cyberspace created by the Internet, the unplanned weaving together of several thousand networks comprising hundreds of thousands of computers and some 25 million people. The Internet grew of out of an early computer communications system, ARPAnet, set up by the Department of Defense. Like the infant universe depicted in hyperinflationary cosmology, the Internet has grown so fast in so many places that no one really knows how big it is or how many people use it. But use it they do. The electronic bulletin board on which high-energy physicists post their latest results gets as many as 10,000 hits every day.

Now just plain folks have begun to move into cyberspace. Several commercial bulletin boards give some five million subscribers access to information bases and the contents of magazines and periodicals; they also offer on-line shopping services. The most prominent are America Online (on which *Scientific American* appears), Prodigy, CompuServe, Delphi and GEnie. The stock of America Online has appreciated by 521 percent since the company went public in March 1992. Magazine and newspaper publishers are looking toward this

The Explosive Growth of Networking

The population boom on the Internet is symbolic of the interest people are taking in networking. Lee Sproull and Sara Kiesler [see "Computers, Networks and Work," page 128] recently noted that since their article appeared, two technology trends have become more important in the way networking can influence the structure of work. They write:

"The first is the explosive growth of connections to the Internet. In 1991 most organizations used networks only for internal communication. In 1995 organizations and their employees find it as easy to access the global electronic world as their own internal electronic colleagues. Through membership in electronic discussion groups, increasing numbers of employees participate in daily discussions with people who share common interests but do not share a common employer. This evolution challenges managers to balance the benefits of employee development and new sources of potentially useful information with the risks of the loss of proprietary information or even the loss of employees as they learn of better job opportunities through the net.

"The second trend is the significant growth in the use of network-based video communication. The growth is likely to increase as video comes to the desktop computer. The ability of employees to rapidly convene ad hoc video meetings will surely be important. Managers and organizations will face significant challenges in learning how to manage video-group dynamics."

future, too. Through the commercial services or independently, Knight-Ridder, the New York Times, Time-Warner, Newsweek and Ziff-Davis have begun to experiment with electronic editions and E-mail communication with readers.

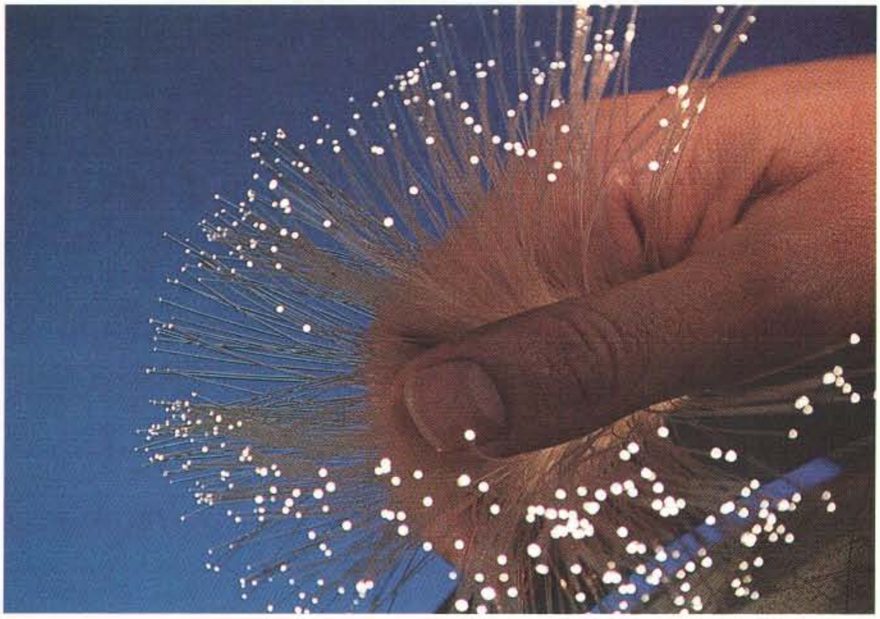
When "Communications, Computers and Networks" was originally published, a software agent called a "knowbot" was being talked about. This useful software agent would serve as an electronic secretary and native guide in cyberspace. Knowing your habits and wishes, it would act without explicit instructions to organize data, hunt for information, protect confidential information and disable viruses that infect your computer. Today several primitive first versions are about to reach the market.

What, then, lies beyond these developments? The view from the huge footprint reveals something of the future. The fusion of computing and communications has already accelerated the transformation of the U.S. economy from an industrial into a service one. According to some estimates, as much as three quarters of the gross national product now consists of services. That percentage is unlikely to change much more. Instead the percentage of value added represented by information technology will increase in all economic sectors. Already the culinary ingredients of a slice of pizza account for a smaller percentage of its cost than do marketing, administrative overhead and legal fees.

Corporate structures will become flatter as information technology enables more decisions to be made at the operating level [see "Computers, Networks and the Corporation," by Thomas W. Malone and John F. Rockart, page 140]. Virtual reality may take the form of the virtual office: computer networks will enable many people in the service economy to do their jobs at home [see "The Computer for the 21st Century," by Mark Weiser, page 78] and communicate with one another by electronically carried voice, sight and word. Ideally, as Mark Weiser points out, computers will be ubiquitous and integrated so seamlessly into daily life that one will seldom take notice of their role.

In the hands of small, nimble firms, powerful computers networked and free-standing may dissolve the economies of scale that normally attach to large organizations. Nicholas P. Negroponte predicts in his article [see "Products and Services for Computer Networks," page 102] that the airwaves may well become a primary carrier for important business and government information, whereas entertainment and consumer traffic will move over the fiber-optic network. On almost any day the business page of your newspaper carries a story or two showing that this trend is getting under way.

The new technology will enable everyone to be a super-couch potato, renting motion pictures and videotapes over the cable system or the telephone line and shopping electronically. (A shopper can now scan the catalogues of more than a dozen other mail-order houses on a CD-ROM.) Some systems enable the buyer to order through the cable channel. Cyber-



SPEEDED TRANSMISSION of communication between computers is achieved with optical fiber carrying lightwaves. The fiber, capable of moving billions of bits per second, is gradually replacing the traditional copper-wire telephone lines that have a maximum capacity of about one million bits per second.

The Incredible Pace of Change

As a senator, Al Gore was instrumental in keeping Congress alert to the communications revolution [see "Infrastructure for the Global Village," page 156]. As vice president, he has continued his interest and activity in this area. He recently outlined the Clinton administration's views on the subject:

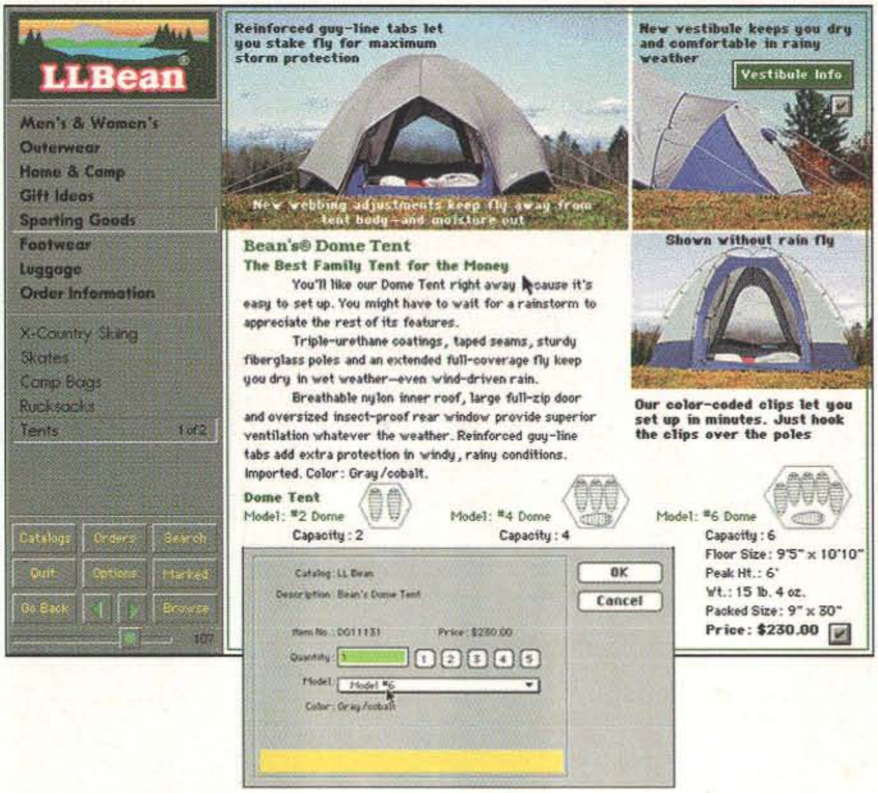
"What we have been seeing with this incredible pace of change, especially in the past decade, really is amazing. But even this change is nothing compared to what will happen in the decade ahead.

"The word 'revolution' by no means overstates the case.

"Our new ways of communicating after this revolution will entertain as well as inform. More important, they will educate, promote democracy and save lives. And in the process, they will create a lot of new jobs. The impact on America's businesses will not be limited just to those who are in the information business. Virtually every business will find it possible to use these new tools to become more competitive. And by taking the lead in quickly employing these new information technologies, America's businesses will gain enormous advantages in the worldwide marketplace.

"The administration's plan for confronting the most pressing telecommunications issues is based on five principles. We will:

Encourage private investment.
Provide and protect competition.
Provide open access to the network.
Take action to avoid creating a society of information 'haves' and 'have nots.'
Encourage flexible and responsive governmental action."



Reinforced guy-line tabs let you stake fly for maximum storm protection

New vestibule keeps you dry and comfortable in rainy weather

New webbing adjustments keep fly away from tent body—and moisture out

Shown without rain fly

Bean's® Dome Tent

The Best Family Tent for the Money

You'll like our Dome Tent right away because it's easy to set up. You might have to wait for a rainstorm to appreciate the rest of its features.

Triple-urethane coatings, taped seams, sturdy fiberglass poles and an extended full-coverage fly keep you dry in wet weather—even wind-driven rain.

Breathable nylon inner roof, large full-zip door and oversized insect-proof rear window provide superior ventilation whatever the weather. Reinforced guy-line tabs add extra protection in windy, rainy conditions. Imported. Color: Gray/cobalt.

Our color-coded clips let you set up in minutes. Just hook the clips over the poles

Dome Tent

Model: #2 Dome
Capacity: 2

Model: #4 Dome
Capacity: 4

Model: #6 Dome
Capacity: 6
Floor Size: 9'5" x 10'10"
Peak Ht.: 6'
Wt.: 15 lb. 4 oz.
Packed Size: 9" x 30"
Price: $230.00

Catalog: LL Bean
Description: Bean's Dome Tent
Item No.: 0011131
Price: $230.00
Quantity: 1
Model: Model #6
Color: Gray/cobalt

SHOPPING BY COMPUTER is now possible through catalogues that can be viewed on an interactive CD-ROM. The example shown is from the L. L. Bean catalogue.

space is sending a wake-up call to the creativity of marketers and advertisers. They will have to move beyond the sense-jarring clamor of 15- and 30-second spots to the art of talking to a consumer who is already in the act of buying—whether the item is a new automobile or a financial product.

But there is a dark side to all this. Like any major social change, the advantages of the communications revolution come at the price of some economic and social disruption. The flattening of the corporate pyramid has already cast out hundreds of thousands of middle managers and highly paid workers, leaving them unemployed or in lower places on the economic scale. In the past decade, the Fortune 500 companies have cut employment by four million—more than 25 percent of the workforce they had when the flensing started. Those workers and their blue-collar counterparts who have seen manufacturing jobs evaporate constitute a large body of frightened, angry, economically disenfranchised and socially declassed people.

Computing and communications technology create a job market that avidly buys intellectual skill. So the entry-level jobs in businesses through which people have traditionally found their way into the economy and eventually to the middle class are fewer, and the ones that remain often lead nowhere. The new technology could contribute to the genesis of Third World countries in the centers of our cities.

The same gap between the technologically sophisticated and the technologically backward may open in the international economy. Some countries, such as Japan, India and many of the nations of the Pacific rim, are already surfing on the demand for hardware and software (at the expense, some allege, of the U.S. job market). But developing countries that cannot muster intellectual resources will find the climb to prosperity very difficult.

And problems abound for individuals who are moving into the timeless, spaceless reaches of the new electronic and photonic territories [see "Civil Liberties in Cyberspace," by Mitchell Kapor, page 174]. In cyberspace, traditional concepts of intellectual freedom and the right to property tangle in conflict. Does browsing in someone else's computer constitute exploration or burglary? Is there a technological fix that can reconcile an individual's right to privacy or a corporation's right to security with the government's right to investigate illegal activity? How can an author or publisher (of software or entertainment) be protected against the prospect that a work may be copied hundreds of thousands of times at a keystroke or deftly reengineered so that the original cannot be recognized? Crosscutting all these issues is the fact that the Internet's informal, helpful communal spirit riles easily at attempts to market products and services in cyberspace. As

The Law and Cyberspace

Anne W. Branscomb [see "Common Law for the Electronic Frontier," page 160] invited comment on her article from lawyers on the computer network Lexis Counsel Connect. One came from David Post of the Georgetown University Law Center, who discussed copyright protection for electronic writing. "Although the jury is still out—indeed, it has hardly even received its charge yet—we are beginning to see that perhaps the very technology that is creating this problem also may be instrumental in solving it. Encryption technology, for example—if its development is not stifled by the administration's ill-advised 'Clipper chip' initiative—will likely allow authors to control at least the first level of distribution by giving them means to restrict access to 'authorized' users. Similarly, 'rendering' or 'viewing' software can provide additional protection by denying access if the user enters unauthorized identification or if the work is not an authorized copy."

Benjamin Wright, a practicing lawyer, discussed electronic contracts: "In the field of business law, we will see traditional courtroom lawsuits over electronic contracts just as we regularly see lawsuits over paper contracts."

David Johnson of Wilmer, Cutler & Pickering in Washington, D.C., commented on the enforcement of network rules: "There are signs that the users of the global network are taking the regulation of this new sphere of activity into their own hands."

Anne W. Branscomb writes in her article [see "Common Law for the Electronic Frontier," page 160], "Cyberspace is a frontier where territorial rights are being established and electronic environments are being differentiated in much the way the Western frontier was pushed back by voyageurs, pioneers, miners and cattlemen."

Should the government be involved in promoting and guiding the development of the information infrastructure the way it was in the creation of the highway system [see "Infrastructure for the Global Village," by Al Gore, page 156]? Congress passed a bill in 1991 establishing a National Research and Education Network that the farsighted senator from Tennessee, Al Gore, proposed before he left the Senate to become vice president. But today what can such intervention add to the exuberant explosion of services whose ground zero is the private sector? Could an NREN program even hinder the process instead of helping it? Some legislators at the state and federal level—together with Vice President Gore and Reed E. Hundt, chairman of the FCC—have argued that telephone companies should be forced to provide advanced communication services to urban ghetto neighborhoods. If such laws are passed, will they succeed in enfranchising citizens of America's Third World, or will they merely raise the cost of moving into cyberspace?

How will the network, and the devices that offer access to it, affect the traditional publishing industries? Will a student want a printed textbook when an electronic text available on-line is updated as a field advances and is illustrated with moving images? In many fields, probably not. Is there a place for a publisher when a writer can offer his or her work directly to the reading public? Yes, presuming that the publisher can enhance communication through good editing and can provide deft marketing that gets the author in touch with a public that wants to read the work.

Magazines and newspapers that publish electronically will be able to make their contents available in real time, updating stories as events unfold. Readers will browse the past as well as the present and, at the click of a mouse, be able to look at an event from many editorial points of view. A news analysis can be accompanied by a politician's press conference or video images of the flight from Kigali. Steffi Graf's backhand can be analyzed by an astute tennis coach, and the virtual Graf on the screen could be made to hit the stroke in a different way [see "When Is Seeing Believing?" by William J. Mitchell, page 110].

PHOTOGRAPH MODIFICATION by computer could be used by a tennis coach or apprentice to analyze the backhand stroke of a virtual Steffi Graf. Here the real Steffi Graf demonstrates how she normally executes the stroke.

Who likes to read a screen? That depends on the screen. One can imagine a traveler about to leave on a flight to London. She would want a laptop computer that is about the size of a magazine. It is equipped with a high-definition display, whose resolution matches that of a fine-printed page. Before departing, the traveler downloads the current issue of *The Economist* and a few stories from that morning's *New York Times* she did not have time to read at home. At JFK, she realizes she has forgotten to bring a book along. So she accesses *Books Out of Print* over the computer's wireless link and picks up a copy of *Those Barren Leaves.* Who knows what other good ideas Huxley may have buried there?

Is Cryptography Doomed?

Gilles Brassard and Artur K. Ekert make the point in their article [see "Quantum Cryptography," page 164] that public-key cryptographic systems have not been proven to be secure. One of them, Brassard recently noted, was broken last April by Derek Atkins, a student at the Massachusetts Institute of Technology, with the help of Arjen Lenstra of Bellcore and 600 volunteers on the Internet. According to Brassard, Peter Schor of AT&T Laboratories has discovered an efficient method for breaking this and other ciphers on the quantum computer. Although this device is "out of reach of current technology," Brassard writes, "if a quantum computer is ever built, quantum mechanics will doom most of contemporary cryptography but will also provide the alternative that we describe in our article."

Networked Computing in the 1990s

Computers began as cumbersome machines served by a technical elite and evolved into desktop tools that obeyed the individual. The next generation will collaborate actively with the user

by Lawrence G. Tesler

In the 1970 film *Colossus: The Forbin Project,* a computer climbs to world domination from the caverns of a hollowed-out mountain, where cabinets decked with flashing lights and spinning tape drives are tended by an army of white-coated programmers.

The stereotype changed between 1977 and 1982, when personal computers entered homes and offices. The new reality was reflected in movies soon afterward: the 1987 film *Wall Street,* for example, portrayed a machine that had shrunk to desktop size, its lights and tapes replaced by a cathode-ray tube, a keyboard and a mouse (that is, a roller that manipulates a pointer on the screen). The priesthood of programmers had vanished, to be replaced by a single user—in this case an investment banker—for whom the computer was a tool, not a calling.

Movies of the 1990s will portray the machines as inconspicuous devices—more like watches than clocks. They will have thin, flat screens, microphones or styli instead of keyboards, and wireless transmitters rather than cable-bound modems. Moreover, their relation to the user will change from that of an isolated productivity tool to that of an active collaborator in the acquisition, use and creation of information, as well as a facilitator of human interaction. Of course, these predictions and others I shall presently make reflect my personal views. Most, however, are widely accepted in the computing profession.

It is generally believed, for example, that the computer will come to play a much more active role by collaborating with the user. The collaborating agent may not even be confined to a particular device. It might move, for example, from palmtop to mainframe data base to desktop. The user, moving from office to car to meetings, might give an electronic agent the following tasks:

- On what date in February did I record a phone conversation with Sam?
- Make me an appointment at a tire shop that is on my way home and is open after 6 P.M.
- Distribute this draft to the rest of the group and let me know when they've read it.
- Whenever a paper is published on fullerene molecules, order a copy for my library.

Programmers will make agents seem intelligent by endowing them with certain reasoning capacities. Agents will consult data bases, and when stumped they will ask the user for guidance. They will be rather like a human assistant, except that they will lack intu-

The Four Paradigms of Computing

	BATCH	TIME-SHARING	DESKTOP	NETWORK
DECADE	1960s	1970s	1980s	1990s
TECHNOLOGY	MEDIUM-SCALE INTEGRATION	LARGE-SCALE INTEGRATION	VERY LARGE SCALE	ULTRA LARGE SCALE
LOCATION	COMPUTER ROOM	TERMINAL ROOM	DESKTOP	MOBILE
USERS	EXPERTS	SPECIALISTS	INDIVIDUALS	GROUPS
USER STATUS	SUBSERVIENCE	DEPENDENCE	INDEPENDENCE	FREEDOM
DATA	ALPHA-NUMERIC	TEXT, VECTOR	FONTS, GRAPHS	SCRIPT, VOICE
OBJECTIVE	CALCULATE	ACCESS	PRESENT	COMMUNICATE
USER ACTIVITY	PUNCH & TRY (SUBMIT)	REMEMBER & TYPE (INTERACT)	SEE & POINT (DRIVE)	ASK & TELL (DELEGATE)
OPERATION	PROCESS	EDIT	LAYOUT	ORCHESTRATE
INTERCONNECT	PERIPHERALS	TERMINALS	DESKTOPS	PALMTOPS
APPLICATIONS	CUSTOM	STANDARD	GENERIC	COMPONENTS
LANGUAGES	COBOL, FORTRAN	PL/I, BASIC	PASCAL, C	OBJECT ORIENTED

Reprinted from the September 1991 issue

ition and the ability to improvise. They will, however, be able to identify recurrent patterns in the user's work, inspect incoming messages and take note of deadlines. Armed with such information, agents will often anticipate needs before the user has expressed them or has even become aware of them. A mobile computer might catch up with its user at breakfast time:

- You asked me when you last recorded a phone conversation with Sam. It was on February 27. Shall I play the recording?
- You scribbled a note last week that your tires were low. I could get you an appointment for tonight.
- Laszlo has discarded the last four drafts you sent him without reading any of them.
- You have requested papers on fullerene research. Shall I order papers on other organic microclusters as well?

Although the commands and responses are represented here as typed English sentences, other forms are possible. They might be handwritten or spoken, more or less grammatical and explicit, in other human languages or perhaps in a more rigid computer language (such as SQL, a language designed for querying data bases). They might be specified by checking off options and filling in blanks on a form or by providing terse answers to a series of questions posed by the agent.

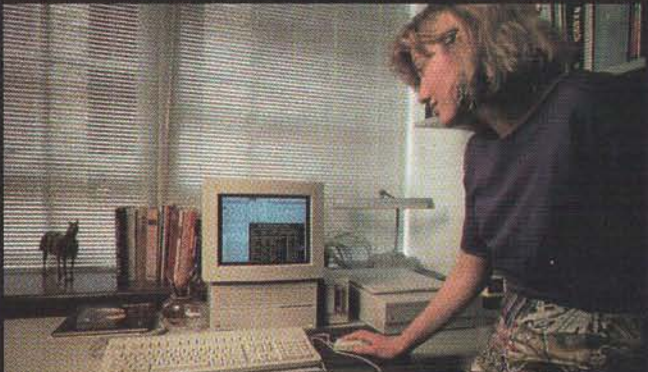
The changes in the computer's role—from cloistered oracle to personal implement to active assistant—have come in distinct waves that can be likened to paradigm shifts, the term by which Thomas S. Kuhn, a philosopher at the Massachusetts Institute of Technology, describes revolutions in scientific thought. Computing paradigms are made possible by steady improvements in a variety of technologies, together with a maturation of the market. They seem to happen at intervals of about a decade.

The original computer paradigm was invented in the late 1940s, when the programmable calculator was designed as an engineering tool; it became commercially practical in the 1950s. The first shift came in the 1960s, when the computer was adopted as a data-processing engine by corporations. The second came in the 1970s, when the computer's services were shared among many subscribers. The third shift, in the 1980s, transformed the computer into a desktop productivity tool for individ-

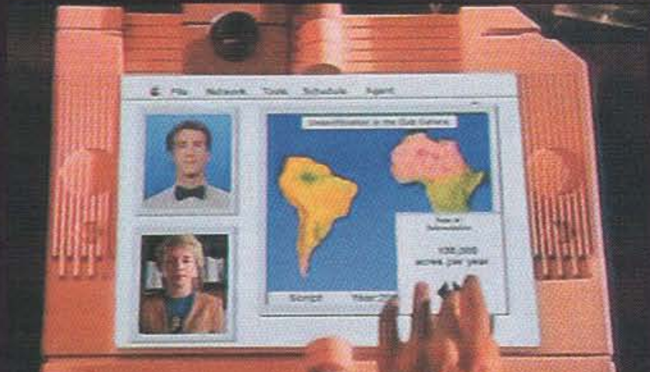
LAWRENCE G. TESLER is vice president of advanced products for Apple Computer Inc. In his former positions with the company, he built a central research group and directed the development of such software products as the applications and user interface for the Lisa computer, the MacApp software development framework and AppleEvents (a mechanism that allows software applications to cooperate in service to the user). In the 1970s Tesler worked at the Xerox Corporation, where he introduced modeless editing and browsing in paned "windows," now a fixture in many interfaces. In the 1960s he managed a small software company, developed easy-to-use animation software and conducted research in cognitive modeling and natural language processing at the Artificial Intelligence Laboratory of Stanford University, his alma mater.

THE IBM 360, circa 1964, processed sequential batches.

THE DECsystem-10, 1975, provided time-sharing.



APPLE COMPUTER'S Macintosh IIci personal computer.

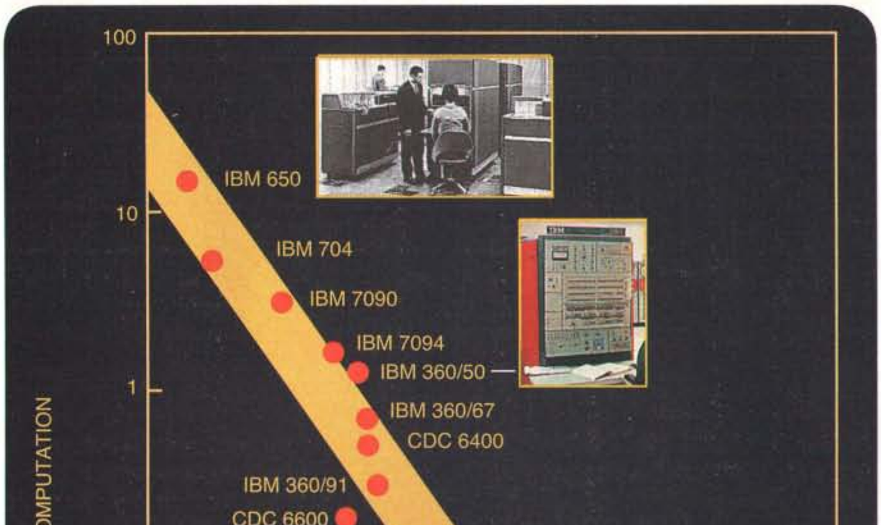


APPLE'S projected software agent: roamer of networks.

uals. The fourth is now under way. Its harbingers are the increasingly networked laptop devices and electronic pocket calendars—mobile machines I call pericomputers. Pericomputers are valuable both for the limited functions they can perform in isolation and for the access they afford to an electronically embodied world of information.

Each shift has fundamentally altered the way people perceive computers. In the 1960s, for example, computing power was dear, and matters had to be arranged at the machine's convenience, as it were. Data had to be processed in huge batches for the technology to pay its way. Only very large organizations could generate such batches. Companies would, for example, run a payroll program to process the week's time cards, then immediately load another "job"—a program with its accompanying data—into the computer. Jobs were coded on punched cards or magnetic tape, and results were delivered as "listings" printed on fanfold, perforated paper. The computer would stop at any error it encountered, in program or data, and return the faulty job to the customer for corrections, a maddening process when reiterated over many hours or even days.

In the 1970s time-sharing made data processing more affordable by allowing many subscribers to split the cost of a computer. This was done by adapting the operating system (by which the computer coordinates its internal activities) so that it divided the machine's attention among such "tasks" as interacting with a particular user terminal. The computer would flit from one task to the next at intervals ranging from less than a tenth of a second to a second or more, depending on the load of work at the time. Computers became easier to use because they could be reached from a terminal and interrogated in real time. That is, one could now work one's way through a problem by asking a question and using the answer to frame the next question.



SOURCE: Victor Peterson, NASA Ames

COMPUTING COST has been halved approximately every three years, as indicated in this graph of the most powerful commercial machines of each era.

Personal computers initially provided services similar to time-sharing, but with much greater convenience. Advances in microprocessors enabled manufacturers to fit a computer on a single chip, making it cheaper to buy a small computer than share a large one. Now that users no longer depended on costly shared facilities, they could work at will on matters that had previously required scheduling. Word processing and chart creation became routine.

As a result, a new class of users was led through a series of increasingly capable software programs, such as spreadsheet applications and page layout aids, that were made practical by the dedication of unprecedented processing power to one individual. Interrogation also became easier because the burden of keeping track of the possible commands shifted from the user to the computer. Now, for example, various choices are displayed in menus and palettes that one can scan visually before pointing out the desired choice.

For the millions of people who could afford computing power for the first time, the benefits were compelling. They no longer had to retype a page to add one sentence, wait weeks for others to provide them with typesetting or layout service, make a decision based on one or two tediously calculated financial projections or present an exciting idea in a drab form. The new services began as a luxury but quickly became de rigueur, much as typewriting had a century before.

It is important to note that the old paradigms do not necessarily die out completely. In this respect, the evolution of computing recalls that of organisms, which often survive in certain ecological niches even though they have been superseded elsewhere by new forms of life. Time-sharing, for instance, still dominates industries that process many transactions—banking, credit-rating services and airline reservations agencies, for example. New applications have also been developed to take advantage of personal computers. Subscribers

“Some eat.

Some get eaten.”

Power Macintosh

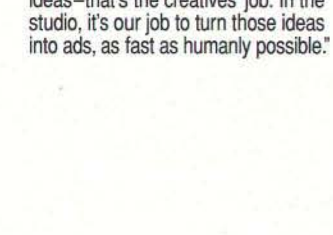
Rick Capps
Capps Studio, Ltd.
Art Studio for Leo Burnett
Chicago

"The Macintosh has definitely helped us grow," says Rick when asked about his studio's expansion from 9 to 115 people in five years. "Of course, producing ads for Burnett clients such as McDonald's, United Airlines, Philip Morris, Oldsmobile, Nintendo and Reebok hasn't exactly hurt us, either."

“We can literally take a concept on a napkin and go all the way to color-corrected digital files and proofs out the door. All on Power Macintosh. All in-house. That not only saves time and increases productivity, it also gives us total control, start to finish. There’s just one problem. Deciding if the mustard stain is part of the concept. Or part of someone’s lunch.”

“Because Power Macintosh is so fast and can push huge files around, we can use it to replace many of the functions of much more expensive, dedicated color correction systems.”

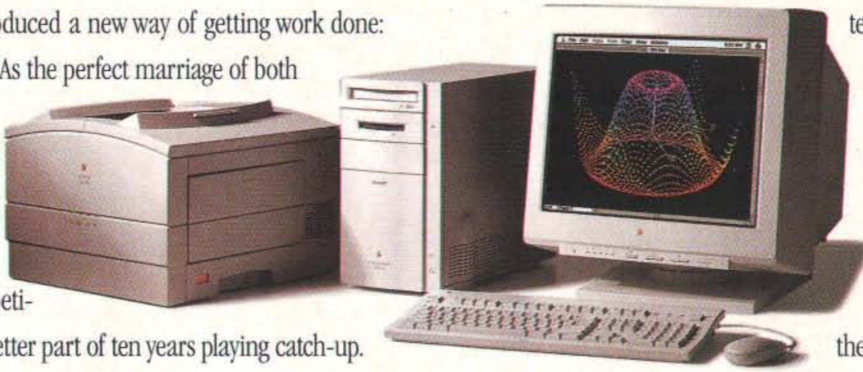
Powered by the RISC-based PowerPC microprocessor, Power Macintosh runs accelerated versions of popular publishing and imaging software at two to six times the speed of non-native versions while providing complete compatibility with existing Macintosh applications.



“Creating big ideas–that’s the creatives’ job. In the studio, it’s our job to turn those ideas into ads, as fast as humanly possible.”

Power Macintosh™ The business Macintosh™ Apple

The next decade of the desktop.



In 1984, Apple introduced a new way of getting work done: the Macintosh® computer. As the perfect marriage of both hardware and software design, the Mac™ quickly redefined the desktop computer, while leaving in its wake a host of competitors that have spent the better part of ten years playing catch-up.

Nonetheless, it's 1995. And it's time to talk about where we are now. And where we're going. Beginning with the Macintosh operating system.

The operating system evolves.

An operating system can be a mere interface to the integrated circuits, capacitors and resistors that comprise a computer and its attendant peripherals. Or it can be an interface to colleagues. To competitors. To the Internet. To entirely new ways of communicating ideas.

Take, for example, System 7.5, the latest incarnation of the Mac OS. With new features like Universal Mailbox, it simplifies communications for people who use on-line services, faxes, voice mail and e-mail. With Apple® Guide, it changes the whole concept of on-line help to one of active assistance. With Macintosh PC Exchange,™ it allows anyone with a Mac to open files created by computers running DOS or Windows. With AppleScript,™ it automates tedious tasks.

And much more. What more, you might ask, could one want from an operating system? Read on.

The invisible application.

Imagine this. You're putting together a research paper for which you need text, graphs, drawings, photos—maybe even a QuickTime® movie—to best make your point. The way things are now, you probably use as many as five different programs to create all those elements—a word processor, spreadsheet, drawing program, scanning software and a video editing package.

OpenDoc

Now imagine this. Instead of working with all those *programs*, you simply work with your *document*. Meaning that you never leave the pages of your research paper to start up a program. Instead, as you need to create or alter text and graphics, your PC automatically calls up the software needed to complete the task.

What you've just imagined is what we and our partners in the CI Labs consortium call OpenDoc.™ What exactly is OpenDoc? It's a new architecture that enables software developers to create small, efficient applications that work together transparently. So you can focus on your work rather than on the computer and the software required to get that work done.

But OpenDoc will do more than simplify the individual user's work. It will also transform today's monolithic applications into smaller, more easily customized programs called "parts." It will create a unified, consistent interface for computer users, regardless of the particular type of computer they use. It will enhance collaboration and sharing across networks and platforms. And, soon, it will be an integral part of the Mac OS.

The universal PC.

Will you have to buy a Macintosh to take advantage of OpenDoc? No. Because, as we've said, the whole point of Power Macintosh™ and innovations like OpenDoc is cross-platform compatibility.

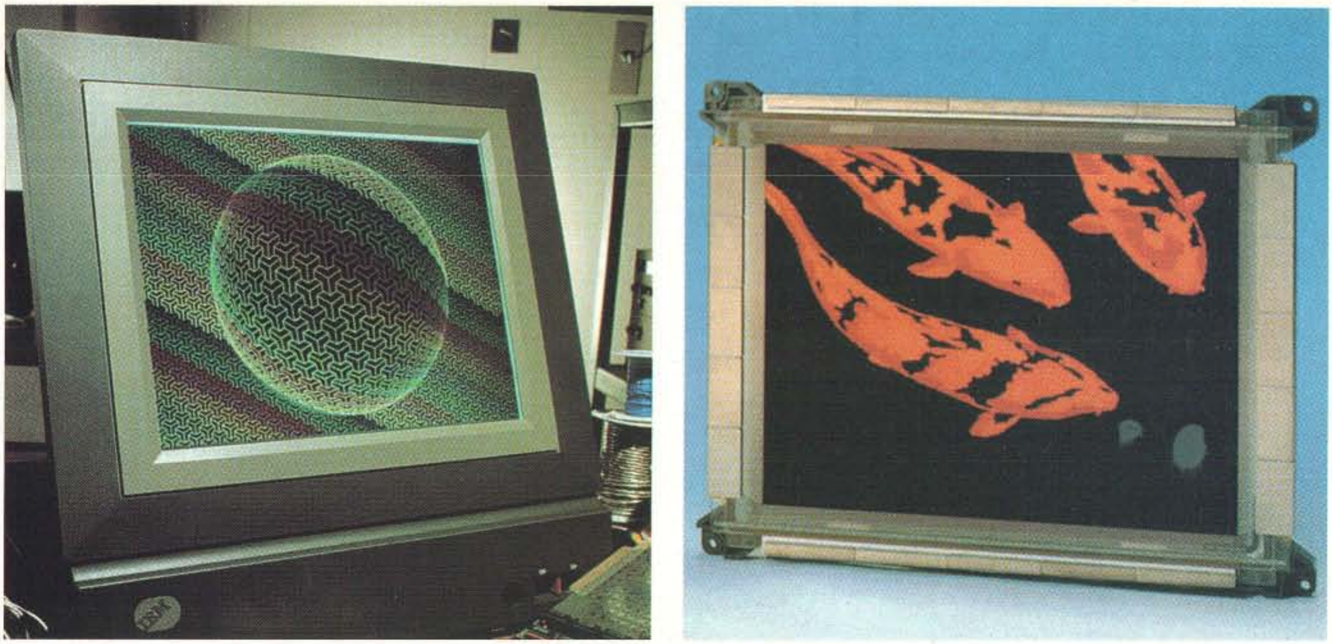
In fact, when Apple, IBM and Motorola's new hardware reference platform for PowerPC™ takes hold, you'll not only be able to use OpenDoc on any PC, you'll also be able to use the Mac OS on any PC. Or, for that matter, OS/2, AIX, even Windows NT. That's because the new platform defines an architecture that can be used by any PC manufacturer to build PowerPC-based computers that will run the OS of your choice.

Taligent	Adobe
	WordPerfect
NOVELL	Lotus
IBM	ci labs

To get more info by fax about these and other new developments, call 800-770-4852. Or contact us via the Internet at http://www.info.apple.com.

Apple

©1995 Apple Computer, Inc. All rights reserved. Apple, the Apple logo, Macintosh and QuickTime are registered trademarks of Apple Computer, Inc. AppleScript, Mac, Macintosh PC Exchange, OpenDoc and Power Macintosh are trademarks of Apple Computer, Inc. PowerPC is a trademark of International Business Machines Corporation, used under license therefrom. All other product names and logos are trademarks or registered trademarks of their respective companies.



FLAT SCREENS, such as this liquid-crystal display of IBM (*left*) and this plasma display of Fujitsu (*right*), are considered crucial for the development of compact, portable computers. Future designs will have to consume less power.

can use their modems to dial on-line news, shopping and banking services.

Similarly, the hardware associated with mobile computing will not eliminate desktop computers. A computer plugged into a wall socket quite literally has more power than one operated from batteries. A person working at a desk will have occasion to exploit the larger screen and more capacious storage of a larger machine, as well as the faster and more accurate input of a full-size keyboard. Another continuing advantage of fixed computers lies in the vast amounts of information that can be fed to them. Wires can carry more data than radio waves, and optical fiber—the projected successor to wire—will widen the gap. This edge will be felt in desktop conferencing and other data-hungry video applications.

Every era of computing reflects the contemporary speed of computation, the variety of information that can be processed, the ways computers can be linked into networks, the software by which such networks are exploited and the ways in which humans can interact with the devices. Speed, the most basic element, has improved largely through miniaturization.

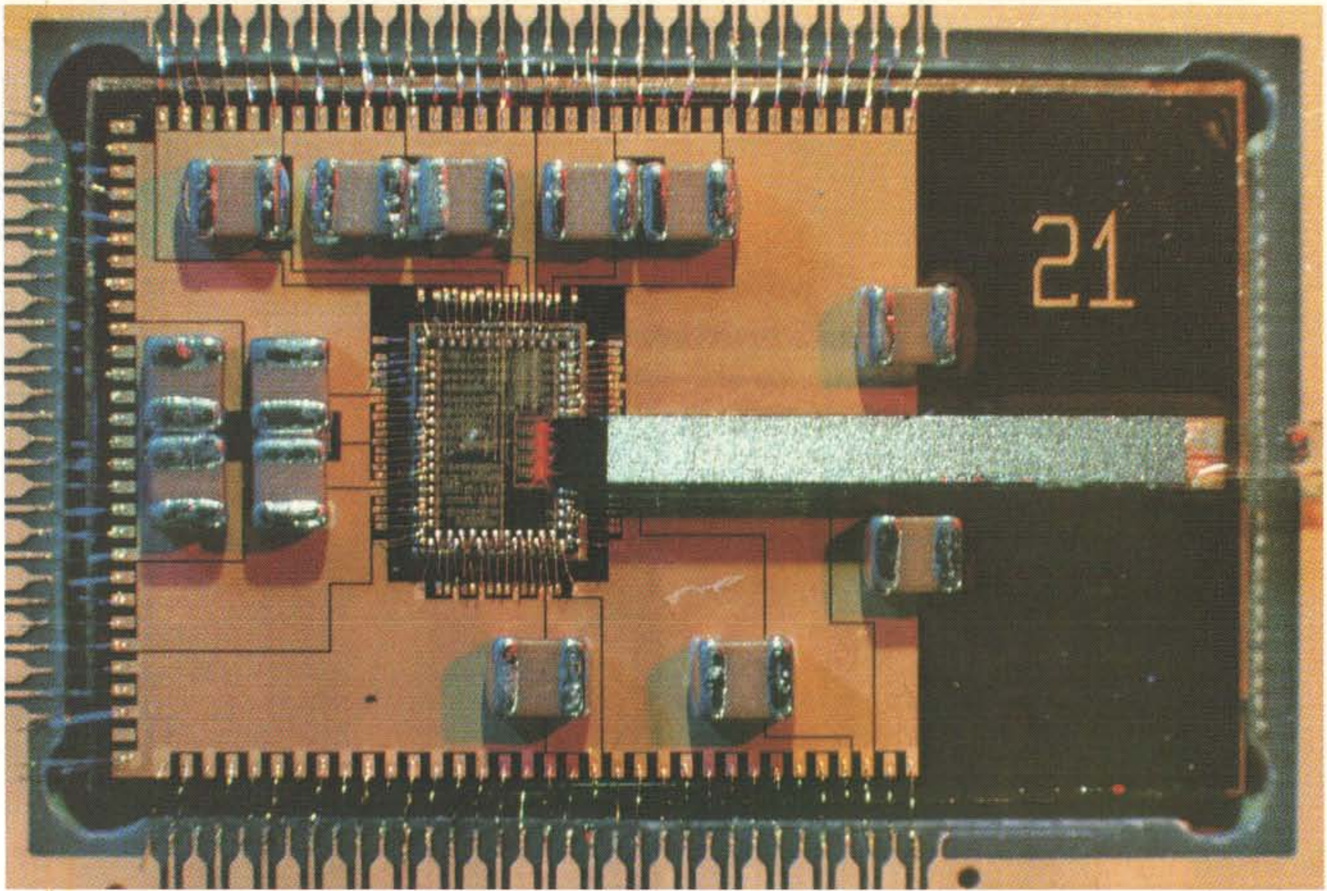
As electronic devices shrink, signals traverse them faster, more operations can be performed in a given period and the cost of computing falls. At the same time, more devices are packed on a chip and thus manufactured in larger batches, reducing the unit cost of manufacturing. Lower costs increase sales, producing additional economies of scale. From the user's point of view, the decline in size yields rapid improvements in usefulness and convenience. Moreover, there is every reason to expect the trend to continue [*see illustration on page 12*].

Thirty years ago computers were assembled painstakingly out of thousands of transistors and capacitors. Today factories mass-produce computers out of a handful of integrated-circuit chips, each comprising millions of components. In the 1990s, instead of packaging each chip in its own centipedelike, flat package, the semiconductor industry will achieve still higher densities of circuits by producing multichip modules in two and even three dimensions. Stacks of chips will form blocks the size of sugar cubes.

Miniaturization also improves other components of computing systems. Memory, for example, was bulkier than paper as recently as 20 years ago. To store in immediately accessible form a million characters of text (about as many as appear in two copies of this issue of *Scientific American*), one then required a disk pack the size of a birthday cake. By the 1980s that many data could be stored on a diskette that fitted into a shirt pocket. In the 1990s it will be carried on a semiconductor device no larger than a credit card.

New kinds of data become usable with each new paradigm. Early batch-processing systems could manipulate only numbers and uppercase letters. Time-sharing systems made possible interactive document and message editing and often supported vector graphics, which render line drawings on the screen. With personal computers came raster graphics, matrices of dots like the pixels that represent television images. Desktop personal computers already support such media as audio, video, animation and three-dimensional graphics, once the domain of expensive workstations. The coming generation of mobile computers will emphasize facsimile and speech, media that are particularly useful to those who are away from the office.

Networks became important as succeeding paradigms increased the number of computers in service. When batch processors were few, there was little reason to try to make them work in concert. Time-sharing, however, is unthinkable without networks: dedicated or dial-up lines must connect numerous terminals to a host computer, either a corporate mainframe or a departmental minicomputer. The obvious next step was to establish wider networks, among many bank branches, for instance, to make possible the electronic exchange of data. Time-sharing users shared centrally located mass storage facilities and high-speed printers connected to the host. They could post electronic mail messages for other users to see when they next logged in at their terminals. The early terminals were "dumb," that is, they served as mere data links. Lat-



IBM'S EXPERIMENTAL NEXIS CHIP converts signals from optical to electronic form fast enough for computers to exchange data in real time over long distances. Such capabilities will be necessary for advanced networks. The chip integrates four gallium arsenide detectors with the circuitry that amplifies and preprocesses their output signal. The fiber-optic channels (*center*) are made visible by the transmission of red light; infrared light is normally used to minimize energy losses.

er versions incorporated minimal computing power, thus pointing the way to personal computers.

In the desktop computing era, computer networks have assumed new forms. Local-area networks connect personal computers to one another and to shared machines—both general-purpose computers, called hosts, and special-purpose machines, called servers, which provide communal files, high-quality printing and institution-wide electronic mail. Wide-area networks interconnect the various fragments of an enterprise, linking the big machines and file servers to desktop machines. As computers become mobile, wireless connections will proliferate until the information network becomes virtually ubiquitous.

Software applications will also change as mobile computers begin to appear in—or out of—briefcases, purses and pockets. Today's portable machines mainly continue the menu style of interaction, but in the next decade supplementary styles will arise. Transcription software, for example, has already appeared that enables portable machines to display printed versions of handwritten words and polished renderings of roughly drawn figures. Related software recognizes hand-drawn symbols that each represent an entire command or idea. Speech-recognition software that obeys spoken commands is also finding wider application as the technology improves.

Most portable computers now run software written for larger machines, but applications packages are being designed for people who are away from their desks. In such situations, users make notes and sketches but lack the time to compose and edit their jottings. They also exchange information with clients and colleagues, check in with the home office and plan their appointments and how to get to them.

Palmtop machines that perform these functions constitute the first full-fledged pericomputers. So far they cannot compute much, but they suffice for situations into which full-function computers cannot go. One can, for example, work out the day's itinerary on a desktop machine and transfer it to a pericomputer for quick reference. Notes taken in the field can be readied for desktop editing; a pocket calendar and a group calendar can be kept in agreement. Such modest applications will establish a customer base large enough to justify writing a wide range of specialized software. New software will then widen the market.

Software, indeed, will change more than any other element in the computing paradigm. For the first time, it will be written with an eye to the group as well as to the individual. The change reflects the widening of networks, which bring users closer to the work and to one another. It is sometimes hard to remember that until the late 1970s computers were operated by technical experts. They were the trained chefs of the industry: users placed their orders and awaited service. The arrival of personal computers enabled people of all backgrounds to use computers. As in home cooking, more effort was often necessary, but the results, if not always quite up to commercial standards, were at least fully under one's control.

In five or six years' time it will be as natural to collaborate through a network as it is to prepare a holiday feast with friends in a common kitchen. The software that supports group activity is called groupware; the collaborative activity itself has been dubbed Computer-Supported Cooperative Work. CSCW can be conducted in one place or many, at the same or different times: a meeting in a conference room, a video conference involving distant sites, an electronic bulletin board through which workers on different shifts cooperate or an electronic mail system through which an author and an editor exchange drafts.

When the calendars of the members of an organization are kept on a network server, meeting times convenient to them all can be found automatically by groupware such as Meeting Maker, a product designed by ON Technology in Cambridge, Mass. The initiator selects a time and an available room, types in an agenda and transmits it to prospective attendees. The software invites the people, asks them whether they plan to attend and gives them the opportunity to comment. The initiator of the meeting can reschedule the meeting with ease, literally by dragging a visual representation of the event from one part of the calendar to another with a hand-operated mouse or stylus. The system then notifies the participants of the change.

Whereas Meeting Maker enhances the productivity of a group, other groupware goes further by supporting abstract intellectual collaboration. Engineers in different parts of the world, for example, will be able to design together as if they were standing in front of the same chalkboard. Networked computers can simulate many existing collaboration tools and overcome some of their limitations. A computer can keep an accurate record of a dialogue for later reorganization, editing and distribution. It can ease the construction, viewing, modification and presentation of models in three or more dimensions. It can transport drafts by electronic mail so that each participant can quickly make editing changes. It can keep track of who changed what and merge nonconflicting changes automatically.

When everyone in a brainstorming meeting has a computer and all the computers are on a network, ideas can be captured and displayed in a shared space for all to see, either on their separate computers or on a wall-projected display. Experience at the Xerox Palo Alto Research Center and other CSCW research centers has shown that people judge ideas that appear on a screen more for their value than by the rank of the contributor. An added advantage is that people need not wait their turn or even speak up to be heard [see "Computers, Networks and Work," by Lee Sproull and Sara Kiesler, page 128].

Groupware becomes more worthwhile when wireless communications allows users to gain access to it at will. Such access is crucial: people want to form impromptu working groups in any room, whether or not it has network sockets, and they most need access to group calendars when they are away from their desks. When this becomes practical, other needs will become apparent, too. A person on the move or in a meeting cannot stare at a screen or issue detailed instructions, for example, but it is easy to give a brief command to an infinitely patient and obedient software agent.

Groupware and other network software will in many cases succeed or fail on the basis of how many tenths of a second it takes to respond to a user's query. In many cases, limited products will beat sophisticated ones by getting the job done immediately. This will be especially true of active services, such as those provided by software agents, that guide a person's decisions in real time. A driver who wants to change course to avoid a traffic snarl, for example, cannot wait long for directions.

It is not enough that software be quick to please; it must also be easy to obtain. The marketing trend can already be discerned. Personal desktop computing stimulated a market for generic applications with mass-market appeal, with the result that software publishers now routinely distribute their products on "floppy" disks or pocket-size diskettes through retail outlets and by mail order. When software distribution becomes electronic, people will be able to order a product and load it into their computers in a matter of minutes. There will be less temptation to make illegal copies, which many people do to save not money but time.

Electronic software distribution will also lead vendors to package software into smaller units that can be transmitted more easily. Consumers will compose their own applications out of separately acquired software, analogous to the components of a home audio system. The advent of such "component" software will reverse the past decade's trend toward ever larger, more expensive and harder-to-learn applications. Instead of upgrading from last year's word processor with 20 features to this year's version with 40, a customer can select desired features from a catalog. Those who need a new feature will be able to receive it immediately and, in some cases, automatically. Subscribers to the Prodigy service network already do so: their personal computers receive updated software immediately after the subscriber logs onto the service.

Many hurdles must be cleared before the promise of ubiquitous networks can be fully realized. Several hurdles may constitute serious bottlenecks, a problem familiar in the early stages of previous paradigms. Although aspects of personal computing had been predicted many years beforehand, for example, it took technologies such as the single-chip microprocessor and semiconductor DRAM (dynamic random-access memory) as well as many commercial factors to fulfill the predictions by 1980. For mobile networked computing to become mainstream in the 1990s, a number of technologies will have to advance in terms of their speed, weight, size, ruggedness, cost and consumption of electric power.

Compact, power-thrifty video screens are an absolute necessity for pericomputers. Liquid-crystal displays (LCDs) are the current standard, although they are only now beginning to match cathode-ray tubes in quality. One approach that promises to solve the problem, pioneered independently by RCA Laboratories and by Westinghouse Laboratories, switches liquid-crystal elements so that they polarize light, thus directing it to the proper color filter. Color LCDs and their fluorescent backlights, however, place a major load on the portable battery. New technologies that emit light directly from the screen may provide alternatives that consume less power. One candidate is the field-emission display, which incorporates thousands of microscopic cathodes.

The power drain of a portable computer's electronics places one of the most important constraints on its performance. The devices that require the least power are therefore attracting the most interest. Among them are chips made according to the process of complementary metal oxide semiconductors (CMOSs); systems that operate at a low voltage (say, 3.3 volts, instead of the standard five volts); circuits that employ few chips, thus lessening the charge on interchip wiring; and architectures that use a slow system clock (which paces all components the way a conductor paces an orchestra).

The two most promising architectures wring more speed from less power in quite different ways. One is called RISC (reduced instruction set computer) because it transfers much work from the hardware to software. Yet most available RISC architectures have been of limited use in mobile machines because they have been designed to optimize performance, not to minimize size and power consumption. One exception is the design of ARM Ltd of Cambridge, England.

The other approach avoids the trade-off between clock speed and power consumption by jettisoning the clock. Such data-flow, or asynchronous, architectures were originally designed to aid in the field of parallel processing. They schedule each processing step to begin only when the necessary data become available. But just-in-time computing introduces great complexities. It is rather like running an orchestra whose members get their cues from one another, rather than from a conductor.

If power is at a premium, one might also try to improve the batteries. The weaknesses of current batteries include the lack of reusability of alkaline designs, the low capacity of nickel-cadmium batteries and the bulk of lead-acid batteries. These weaknesses can sometimes be offset. Miniaturization has improved capacitors, for instance, so that they can accumulate power from low-output sources and supply it in periods of high demand. Software has also been used to drain power automatically from nickel-cadmium batteries before each recharge, thereby averting a permanent loss of storage capacity (or so-called memory effect). New batteries based on nickel-metal hydride and lithium also offer great promise, as do photovoltaic cells, which are already used in calculators. Unfortunately, the conservation of energy makes it impossible for any manageable array of solar cells to supply a backlight bright enough to overcome the sun's glare.

The technologies of handwriting and speech recognition are important because they make mobile computers easier to use. That mission must always be kept in mind; otherwise, one may be tempted to judge a system solely by the number of words it can distinguish or the percentage of errors it makes. Handwriting recognition has been touted for bringing computers into the lives of people who do not like to type. Even those who do like to type cannot do so when standing or holding a machine. In Japan, where the language has a profusion of written symbols that keyboards cannot easily handle, the technology is considered a key selling point.

Experience gained with handwriting recognition will be valuable because it can, to some extent, be transferred to speech recognition. Both technologies require, in the more sophisticated applications, a detailed linguistic analysis by which the software raises its chance of recognizing a word or phrase. Both face the challenge of finding where one word ends and the next one begins. Finally, both must be trainable, so that they can adapt their representations of letters or sounds to the handwriting and pronunciation of the user.

Versatile speech recognition requires substantial computing power, but lesser hardware can provide limited capabilities that often prove useful. It is hard to program any computer to take continuous dictation without mistaking the end of one word for the beginning of the next. But it is easy to program even a hand-held computer to distinguish "yes" from "no." A handful of commands may be all most users need to utter when their hands are not free. It has been shown, for example, that mail sorters become much more productive when they call out zip codes instead of dropping their work to fill out a form. Yet this application requires that the system distinguish only 10 words: those for the numerals.

Recognition systems can identify a stroke of the stylus or a trill of the voice by analyzing it into a set of features. Such features can then be compared against a template stored in the computer's memory. Alternatively, the features can be analyzed by neural networks, so called because they mimic the way in which the brain is believed to work. A computer links a digital representation of neurons into a network, then selectively changes the weights assigned to each connection. A process of trial and error then "trains" the system to distinguish various signals.

Another technical challenge to wireless pericomputers is the capacity of the electromagnetic medium itself. To prevent one transmitter from drowning out another, networks within a given building will group transmitters into cells as small as a room. Larger cells will require radio, although local ones may choose infrared light because it uses a part of the electromagnetic spectrum that is not regulated by federal and international authorities.

The limited carrying capacity of that spectrum was exemplified about 10 years ago, when the vogue for citizens band (CB) radio collapsed of its own

PERSONAL ORGANIZER brings networks to the palm of one's hand. These mobile devices can be downloaded directly (*as shown above*) or by telephone.

MASTER AND MAN: the image of the computer has changed with its mode of use. When the machines were expensive and hard to program, they were portrayed as remote tyrants, as in the 1970 movie *Colossus: The Forbin Project*. In this scene, Dr. Forbin is shown leaning against Colossus, the computer that he designed and by which he is finally mastered. Will tomorrow's computer be able to play the benign role of the ideal personal servant—at once ubiquitous and inconspicuous?

weight when users could no longer make themselves heard above the din. Yet that crunch on band space will seem as nothing once mobile computers start blasting megabytes of data across the ether. Even the adoption of a system of cellular relays, such as the one now used for car telephones, might not defer saturation for more than a few years.

Cellular systems will in any case be necessary because mobile machines have to be tracked as they move. If two collaborating users get out of range, the network will have to reestablish the connection transparently through intermediary routing equipment. Even so, there will be times when mobile devices lose contact. Software must therefore be designed so that people affected can easily resume interrupted communication.

No less important than the changes in technology are their human effects. Every new paradigm has molded the way users perceive their status in relation to the computer. In batch and time-sharing days, many users felt subservient; desktop computers gave them independence; mobile networked computers will bring them freedom.

The networked computer will thus challenge not only business but also society. Universal connectivity raises issues of security and personal and business privacy. It also raises the question of the distribution of power. The chasms between rich and poor could widen, for example, if the latest computing paradigm creates still more opportunities for educated people and still fewer for the uneducated.

For reasons of social equity and economic efficiency, it will become more important than ever to educate all people so that they can benefit equally from the information resources that are about to become available. Should the benefits of networks become general, democracy might well be enhanced. Such a pattern can be discerned in Eastern Europe, where the recent revolutions appear to have been helped along by the presence of personal computers, copiers and facsimile transmitters.

The computer, then, has changed much since the days when the movies portrayed it as a relentless superbrain, extending its tyranny to all the world. Its popular image has also changed, albeit with a slight lag. Yet the day is not distant when mobile computers will be common in movies and movie sets alike, where they will manage everything from the scriptwriter's 23rd draft to the key grip's accumulating overtime pay. In that day, art will catch up to life, and the computer will take on a new persona.

FURTHER READING

COMPUTERS AND COMMUNICATIONS. Koji Kobayashi. MIT Press, 1986.

THE MEDIA LAB: INVENTING THE FUTURE AT MIT. Stewart Brand. Viking Penguin, 1987.

COMPUTER-SUPPORTED COOPERATIVE WORK: A BOOK OF READINGS. Edited by Irene Greif. Morgan Kaufmann Publishers, 1988.

GROUPWARE: COMPUTER SUPPORT FOR BUSINESS TEAMS. Robert Johansen. Free Press, 1988.

IDEAS AND INFORMATION: MANAGING IN A HIGH-TECH WORLD. Arno Penzias. W. W. Norton & Company, 1989.

SHARED MINDS: THE NEW TECHNOLOGIES OF COLLABORATION. Michael Schrage. Random House, 1990.

Communications, Computers and Networks

By fusing computing and communications technologies, we can create an infrastructure that will profoundly reshape our economy and society

by Michael L. Dertouzos

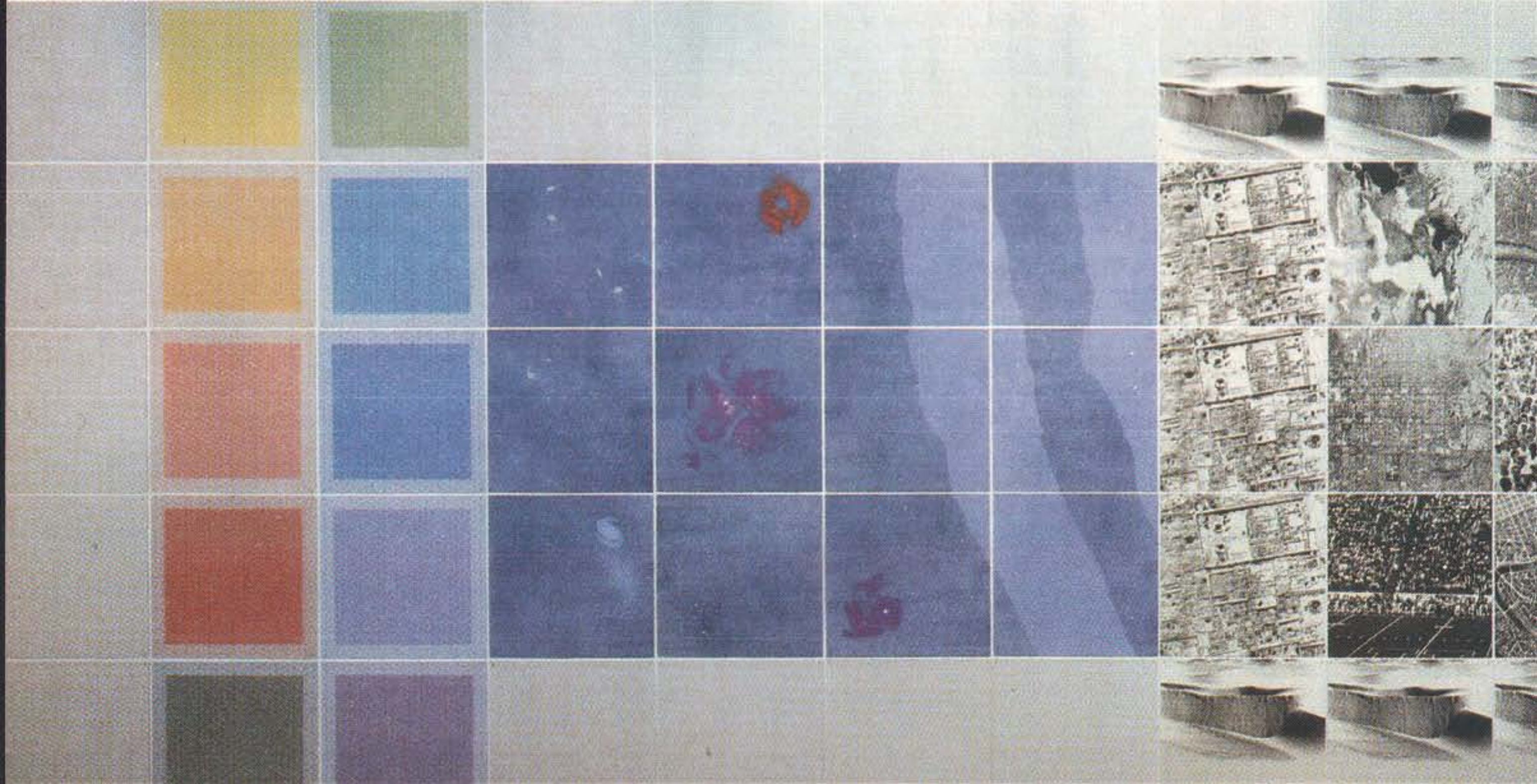
The agricultural age was based on plows and the animals that pulled them; the industrial age, on engines and the fuels that fed them. The information age we are now creating will be based on computers and the networks that interconnect them.

The authors of this issue share a hopeful vision of a future built on an information infrastructure that will enrich our lives by relieving us of mundane tasks, by improving the ways we live, learn and work and by unlocking new personal and social freedoms.

Since the timid debut of the first useful computers almost half a century ago, there has been no shortage of opinion, speculation and prediction about the magical new world they promise. Why, then, devote this entire special issue to this topic at this time? What is new?

A growing opportunity has attained critical mass as a result of a twofold serendipity: dramatic improvements in the cost-performance ratios of computers and of communications technologies. Independent of each other, com-

"WINDOWS," an eight-by-30-foot mural made of enameled steel by Anna Campbell Bliss, explores the worlds being opened by the innovative use of computers. The artist attempts to span time and place, showing how computers let us see phenomena from the microscopic to the macroscopic. The palette of colors (*left*) is typical of graphics programs. Representations



puting and communicating tools have been improving at the annual rate of some 25 percent for at least the past two decades. This relentless compounding of capabilities has transformed a faint promise of synergy into an immense and real potential.

Computers have grown so powerful and cost-effective that they can be found nearly everywhere doing nearly everything: Supercomputers, manipulating billions of commands per second, forecast the weather and analyze complex medical images. Sensory computers respond to spoken sentences and visually recognize parts on assembly lines. Robotic computers turn those parts into full products. Like the telescope and microscope, computers are opening up new realms for scientists by simulating everything from astronomical collisions to molecular reactions. Fifty million personal computers along with thousands of varieties of software packages help people at work and at home. And millions of computers disappear every year into the cars, microwave ovens, telephones and television sets that they control.

At the same time, the reach and speed of networks have increased by equally awesome strides: millions of miles of glass fibers handle most long-haul communications and are capable of relaying data at speeds of up to a billion bits (gigabits) per second. Local-area networks have become indispensable webs, wiring numerous buildings and neighborhoods. Cellular and other wireless networks reach people while they are driving or even walking. And now these two giants, computers and networks, can be fused to form an infrastructure even more promising than the individual technologies.

During the past 10 years, we have, moreover, learned lessons about the many useful possibilities offered by the information infrastructure and the difficulties in building it. For instance, throwing computers and networks together without careful development of common conventions that enable them to communicate easily does not lead far.

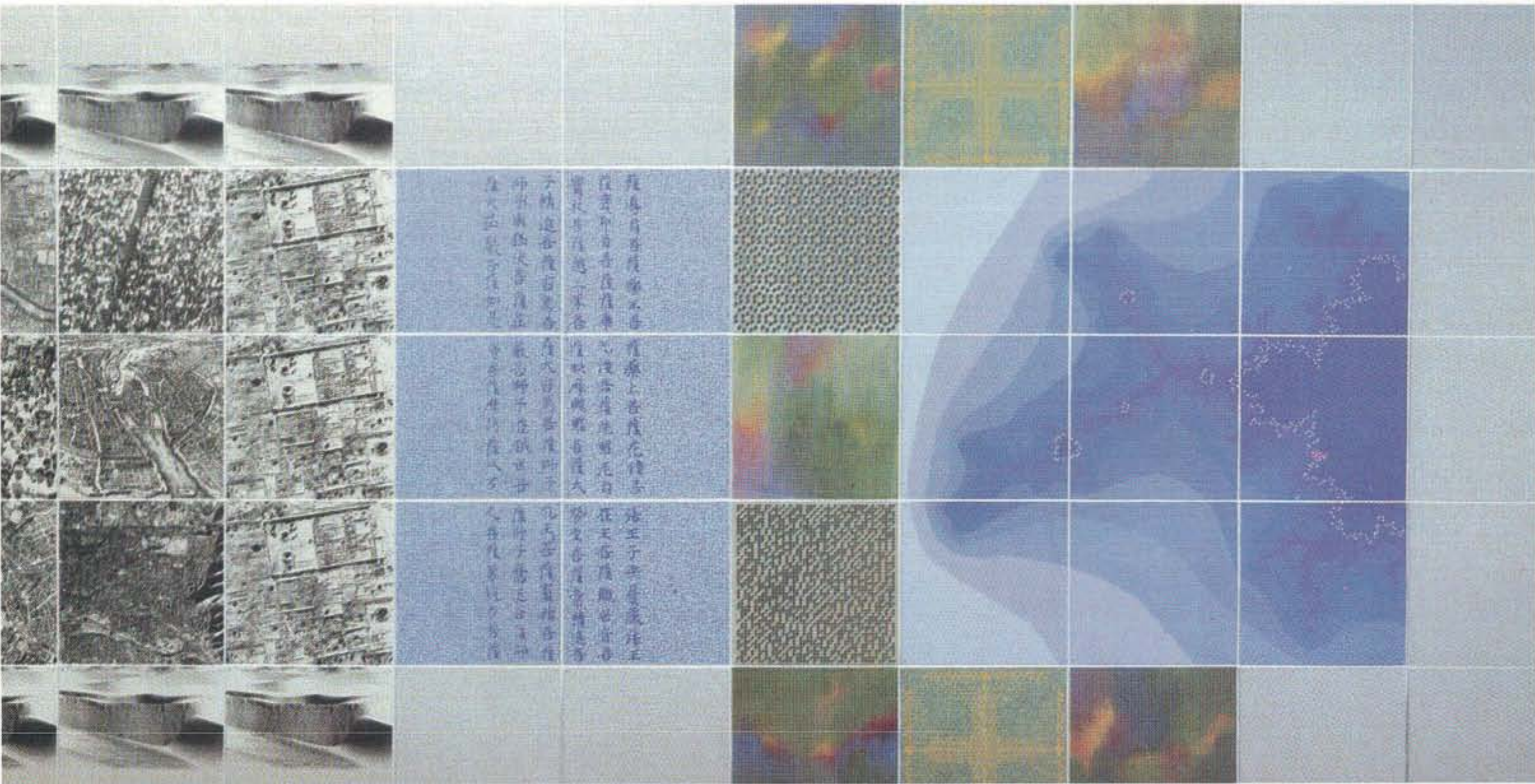
As we lay the bricks of the information age, trying to envision the ultimate edifice and its uses is as challenging for us as it would have been for writers in the late 1700s to anticipate the automobile, the helicopter, the jet airplane and the myriad of other modern engines along with all that we do with these machines. So we simply offer readers our best impressionistic glimpses into this future, its underlying technologies and surrounding issues.

In a world in which hundreds of millions of computers, servants to their users, easily plug into a global information infrastructure, business mail would routinely reach its destination in five seconds instead of five days, dramatically altering the substance of business communications. A company's designers and marketers would actively collaborate on a product, even when located a continent apart and unable to meet at the same time. Consumers would broadcast their needs to suppliers, creating a kind of reverse advertising. Many goods would be ordered and paid for electronically. A parent could deliver work to a physically distant employer while taking care of children at home. A retired engineer in Florida could teach algebra to high school students in New York City. And from a comfortable position in your easy chair, you could enjoy a drive through your next vacation spot, a trip through the Louvre or a high-definition movie rented electronically, chosen from the millions available.

MICHAEL L. DERTOUZOS directs the Laboratory for Computer Science at the Massachusetts Institute of Technology, where he is professor of computer science and electrical engineering. He is a member of the National Academy of Arts and Sciences in Greece and a member of the U.S. National Academy of Engineering. He recently chaired the M.I.T. Commission on Industrial Productivity. He received his Ph.D. from M.I.T. in 1964 and joined the faculty at that time. He enjoys playing with computers and sophisticated electronic gadgets, along with dreaming up new interactions between technology and people.

Yet the information age has its shad-

of outer space are juxtaposed with images from pre-Columbian Peru as well as present-day Salt Lake City. Bliss contrasts Chinese calligraphy with randomly generated alphanumeric printing, followed by an interpretation of computer memory. A "fractalscape" evokes the future. The work is on display at the Data Center located at the Utah State Capitol.



ows as well. Will these new technologies widen the gap between rich and poor? I suspect so. Will they cause us to be inundated with "infojunk," mountains of information irrelevant to us? Yes, but we can also use the technologies to shield us from such perils. Will they threaten to dehumanize people? I doubt it. But will these technologies also increase white-collar crimes and violations of privacy? We don't know; we must be vigilant.

And as has been the case for much technological change, the glorious possibilities we describe in this issue stem more from opportunism than they do from pressing human need. Consequently, we, the designers and users of this information infrastructure, bear a serious responsibility: we must understand the value and role of information so that we may better channel our technological miracles into useful rather than frivolous, if not dangerous, directions.

VOYAGER, an interactive speech-recognition system that provides a street guide to Cambridge, Mass., is being developed at the Massachusetts Institute of Technology.

Information touches all human activity. It comes in a multitude of different shapes—speech, pictures, video, office work, software, great art and kitsch, invoices, music, stock prices, tax returns, orders to attack, love letters, novels and the news. We have also created many ways of conveying information, from cheap, large newsprint pages to postal systems to telephone, radio and television networks. Virtually all these schemes require humans at the receiving end to understand and then to act on the incoming information.

Similarly, computers and networks are bound together by information and can fulfill roles roughly analogous to those of people and their communications schemes. Computers accept, store, process and present information; the networks move information among the machines they interconnect. Computers can manipulate information far faster than people ever will. But unlike people, machines almost never understand the messages they are manipulating. To them, information is only a deceptively uniform sequence of numbers—ones and zeros.

One key idea behind the information infrastructure is to relieve people of a good deal of the work of communicating and processing information. To do so, the machines must intelligently handle some of the dazzling diversity of concepts that information represents. They must therefore begin to understand, even at a crude level, what the ones and zeros mean. In contrast, the current state of affairs is tantamount to a community of telephone callers trying to work with one another using only meaningless grunts of different loudnesses to convey messages.

Understanding the value of information, though, is difficult even for people. Although we are continually besieged by information, we have at best only an intuitive grasp of its meaning and almost no sense of how to value it. How valuable is a 300-page report on a company's stock? What makes a 15-page booklet or a well-placed stock tip more valuable?

Until a better explanation is devised, let me suggest that information has economic value to people only if it can lead them to the acquisition of tangible goods. Similarly, information has intangible value if it can enable them to satisfy less tangible human desires. An encyclopedia publisher, for instance, will find a mailing list of prospective buyers useful because it might increase sales. Watching a soap opera has value for those people who want to experience heartrending emotions.

Because information leads to goods only indirectly, it seems reasonable to value it as a fraction of the worth of the tangible goods to which it leads. If information leads to goods through intermediate pieces of information, each one derived from the others through processing, then all the intermediate data and programs should also be somehow valued backward from the end results.

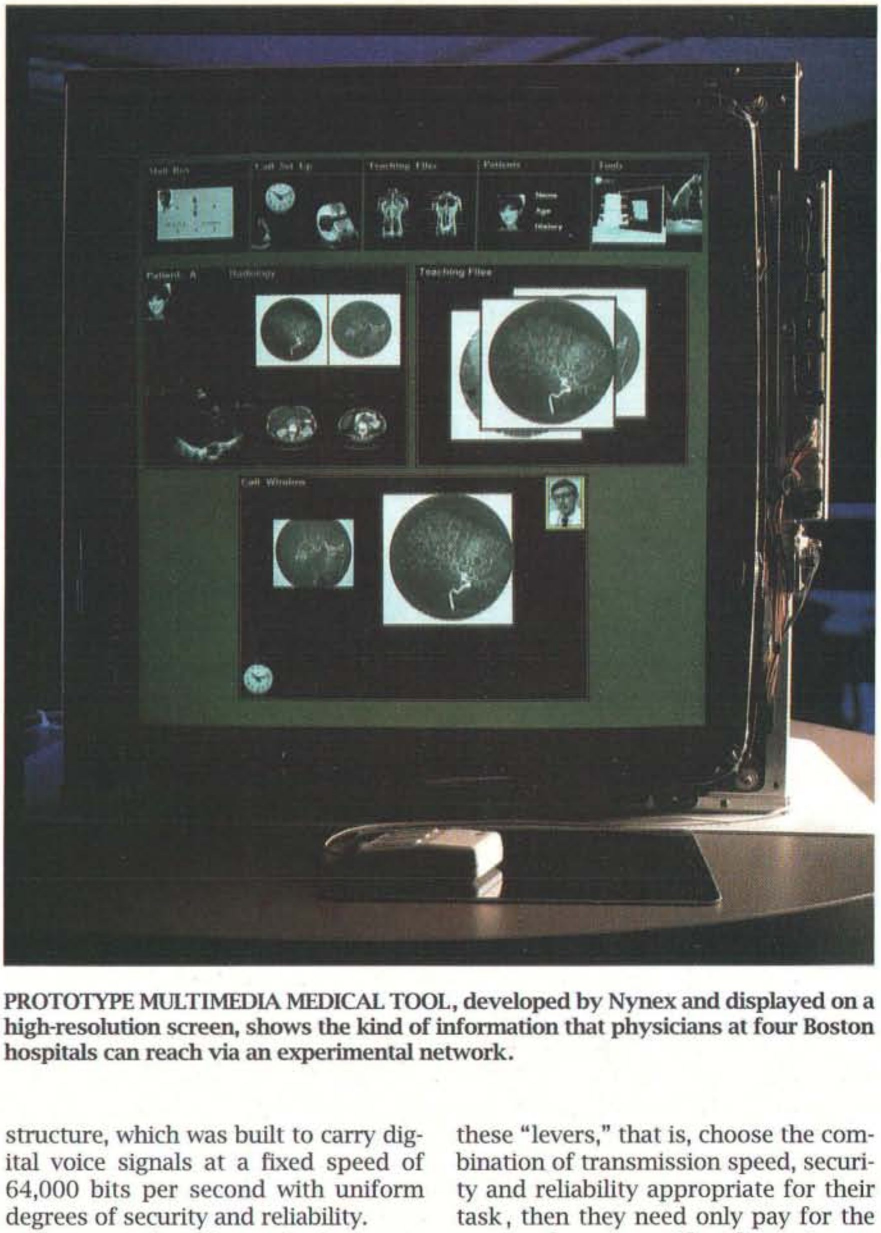
Using these ideas, we can measure the economic value of all that will transpire on tomorrow's sprawling computer-and-network complexes as a fraction of the tangible goods to which they will lead. Today an industrially advanced and wealthy country such as the U.S. places the value of its computer hardware and software, including the work needed to run computer systems within organizations, at almost a tenth of its gross national product—roughly about $500 billion. Yet because some 60 percent of the work force have jobs that involve information, the value of computerized information handling may well grow to an even larger fraction of the U.S. economy.

Valuing information this way points out some sobering lessons to bear in mind as we embark on the information age. To be valuable to us, tomorrow's computers and networks must help us achieve our tangible goals even as they shield us from the barrage of infojunk produced by others aiming to achieve their own goals. This objective raises to a still higher premium the need for computers and networks to understand enough content so they can isolate, simplify and present useful information and reject irrelevant data before they clutter users' lives.

Thinking about the value of information brings home another lesson: that information and its processing have less value in poor countries because there are not many tangible goods to which they can lead. Yet there are other ways in which information can play an important role in these countries. It can help teach people to take better care of their health, to fix machinery or to establish better farming practices. Information might also improve the distribution of food and medicine. Countries with well-educated pools of labor, such as India, may even earn foreign exchange by writing software for overseas consumption.

On balance, however, information more naturally boosts the wealth of those who already have material goods—simply because those communities already have so many tangible goods to which information can lead. Unless wealthy countries see it as their duty to help developing nations make good use of the evolving technologies, the information age will likely widen the rift between the haves and have-nots.

Rich nations must also remember that if they become enamoured of and blinded by the glamour of the information era and neglect to produce and improve tangible wealth—such as food, manufactured goods, natural resources and human services—the information colossus will lead to nothing and so will collapse. Information is, after all, secondary to people's principal needs—food, shelter, health and human relationships. In a crisis, even the most dedicated hackers would trade millions of bytes of software for a few bites of bread.



PROTOTYPE MULTIMEDIA MEDICAL TOOL, developed by Nynex and displayed on a high-resolution screen, shows the kind of information that physicians at four Boston hospitals can reach via an experimental network.

We begin to reap the value of information when we have created an infrastructure that leverages our work. The existing web of computer, telephone, broadcast and other kinds of networks does not constitute the kind of powerful information infrastructure that we envision—no more than the thousands of U.S. dirt roads in the early 1900s made a national highway system. Intrepid travelers could drive from one place to another by navigating over those twisting passages, but such trips were slow and difficult.

Well-established infrastructures, such as the present highway system, the telephone network and the electric power grid, have several simple but powerful properties. They are widely available: accessible to practically every American is a well-paved road, at least one telephone and several electric power outlets. These infrastructures are also easy to use. No more effort is required than pushing a plug into an electric socket or talking into a telephone. And most important, these infrastructures serve as the foundations for countless useful activities. We conduct business deals and family chats over the telephone, ferry people, food and every conceivable good over the highways, and so on.

Using these yardsticks, we can see that no information infrastructure exists anywhere in the world today. Learning how to use a computer still causes headaches. And it is impossible to build even one application that can be used by all the nation's computers.

To escape the present chaos and to fashion our computers and networks into a true information infrastructure, we must endow networks with three key capabilities: flexible information transport capabilities, common services and common communications conventions.

Flexible transport means that the infrastructure can carry information among computers with various degrees of speed, security and reliability. This requirement is vastly different from the capabilities of the telephone infrastructure, which was built to carry digital voice signals at a fixed speed of 64,000 bits per second with uniform degrees of security and reliability.

Humans only talk on telephone networks; computers carry on far more diverse activities. Computers can convey information at a wide range of speeds, from a few thousand bits per second for sending a brief text message to tens of millions of bits (megabits) per second for shipping high-resolution video.

Depending on their tasks, computers also have variable access and security needs: arranging a personal loan electronically demands more security than does chatting openly on an electronic bulletin board about the Boston Celtics. Some messages must also be sent with more precision than others. Conveying software, funds or lifesaving medical data requires perfection. A certain laxness, on the other hand, is tolerated in the transmission of photographs, in which case a few lost bits do not alter the meaning of the message.

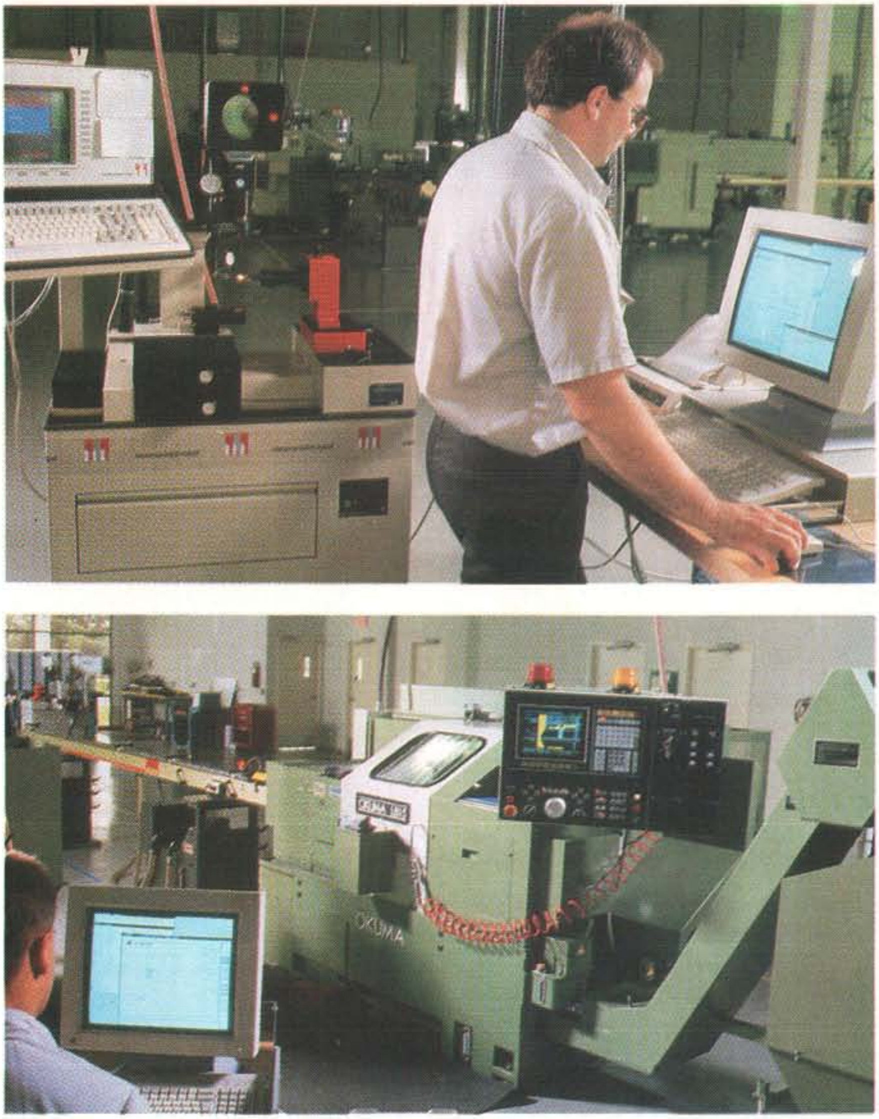
If users or their computers can set these "levers," that is, choose the combination of transmission speed, security and reliability appropriate for their task, then they need only pay for the service they want. The alternative—a highly secure and lightning-fast transport service that would satisfy all potential needs—would be so expensive that it would never become widely used.

As Vinton G. Cerf points out, the U.S. is making gradual progress in building the hardware for a flexible network [see "Networks," page 44]. Telephone companies and other network builders continue to lay down optical fiber lines that are capable of transmitting thousands of times—or eventually hundreds of thousands of times—more traffic than the traditional copper wire. An emerging standard for future telephone service, called broadband integrated-services digital network, will let users transmit data at speeds of up to 150 megabits per second. By the end of the 1990s, the U.S. may have in place a working gigabit network, capable of conveying video images that will rival prize-winning photographs in their vibrancy and crispness.

Flexible service also means that people can tap into the infrastructure wherever and whenever they want. Wireless networks based on cellular and satellite systems will make this access possible by letting automobiles and people walking on the street be part of the world's information infrastructure.

The second component of a proper information infrastructure is a set of common services that would be available to everyone. At a minimum, there must be a few basic and necessary common resources, such as directories—electronic white and yellow pages—of users and services. But there could be richer, universally shared resources as well: government tax codes and regulations, census data, the paintings in the National Gallery and the 15 million books in the Library of Congress.

To understand the third and most important ingredient of an information infrastructure—common communications conventions—consider again the frustrations of people trying to communicate with grunts over the telephone. Any pair trying to communicate may assign meanings to a few specific grunts so they can understand each other. Crude as it may sound, that method is used today within groups of interconnected computers. A company and its suppliers, or even people within one large firm, draw up agreements to establish specific formats for exchanging information. Although this approach works well for a small number, it becomes absurd for a larger community. It is far more economical for everyone to agree on some conventions—in other words, on a common language.



PAPERLESS FACTORY at Raychem-Advanter in Richmond, British Columbia, relies on computers at every step in the manufacture of very small batches of aluminum adapters in thousands of sizes. A technician, who receives on-line information about the next job, uses an optical system to measure the new cutting specifications and so can reset the lathe in just 12 minutes (*top*). As the computerized numeric-control lathe begins cutting the part (*bottom*), a technician can also monitor the work using a continual stream of quality-control data.

One way to create such a language for computers involves what I like to call E-forms, or electronic forms. E-forms, the computer equivalent of mail-order forms, would be instantly recognized by any computer on the network that needs to do so. A handful are already in use. Several years ago corporations agreed on a broad format called the electronic data interchange (EDI) for settling business transactions, including sending invoices and ordering parts. Electronic services, such as Prodigy (developed by Sears, Roebuck and Co. and IBM), CompuServe and Dow Jones News/Retrieval, offer E-forms for ordering airplane tickets and some merchandise. Unfortunately, these early E-forms are not universally used by computers but are restricted to those clients who pay for the services.

Eventually E-forms might be filled in by speaking rather than by typing. A few prototypes exist: bond traders at Shearson Lehman Brothers and about 40 other firms use voice-activated assistants to record sales of government securities. The machines have small vocabularies, however, and must be trained to understand a specific broker. A research system at the Massachusetts Institute of Technology, called the Air Travel Information System (ATIS), helps people wishing to book flights. It understands continuous speech by any speaker—including those who have accents—provided the person wants to order a ticket. People normally phrase questions in many different ways. Yet the computer need only fit the question asked into one of a few categories. As a result, voice E-forms are much more feasible than general-purpose systems for understanding speech.

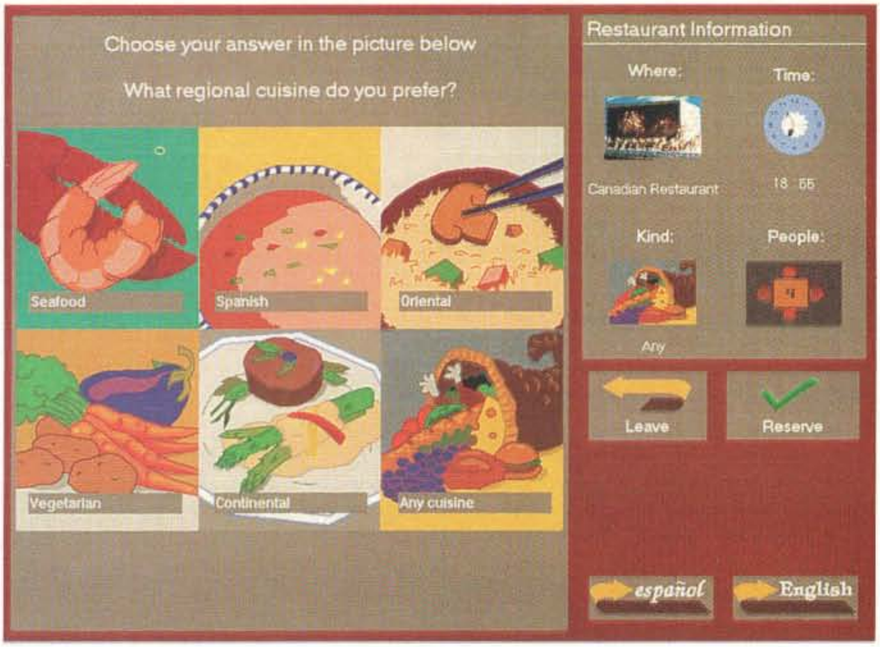
Whether typed or spoken, E-forms can also bridge different languages. For instance, an American ordering a pair of Italian shoes need only fill in an E-form in English; it would then automatically be translated into the corresponding E-form in Italian. Such facile translation may make E-forms a key factor in the European Union's quest for commercial unification, providing the participating countries with an easy means to overcome linguistic barriers in routine business transactions.

An alternative to this E-form Esperanto would be a Knowbot, first developed by Cerf and Robert E. Kahn at the Corporation for National Research Initiatives (CNRI). Knowbots are programs designed by their users to travel through a network, inspecting and understanding similar kinds of information, regardless of the language or form in which they are expressed.

Suppose, for example, that you wanted to create a list of all the available car models that have enough backseat leg room for tall people and cost less than $18,000. If details on various vehicles are available through the information infrastructure but are represented in different formats by the manufacturers, you might unleash a Knowbot to roam the net and scan the various forms. Your Knowbot would understand enough about the different ways the same kind of information may be represented to glean the relevant details from every entry. It would then process and present the information to you in a useful and familiar way. An early prototype Knowbot, designed by the CNRI, combs through data bases developed by the National Library of Medicine, looking for salient facts on people listed in publicly available data bases.

Because Knowbots can be tailored to meet a person's needs and can tolerate diverse representations of information, they look more desirable than E-forms. But as we ask these programs to act with greater intelligence, they become far more challenging to build.

Beyond creating a proper infrastructure with these three elements, we must also develop computer hardware and software that can more naturally connect people and machines to the information infrastructure and, as a result, to one another. Lawrence G. Tesler discusses how designers are trying to transform hardware into tools as accommodating as a wristwatch [see "Networked Computing in the 1990s," page 10]. Most important, we must rethink how to link people and machines at the cognitive level so as to communicate understanding instead of grunts. To this end, Mark Weiser proposes we find ways to make the hardware fade into an inconspicuous feature of the environment, even as it enhances our understanding of the events and people around us [see "The Computer for the 21st Century," page 78].



Choose your answer in the picture below

What regional cuisine do you prefer?

Restaurant Information

Where: Canadian Restaurant

Time: 18 55

Kind: Any

People:

ON-LINE RESTAURANT RESERVATIONS can be made through a computerized guide (developed by IBM) to the 1992 Universal Exposition in Spain. Users pick a cuisine. Networked computers then relay the relevant restaurants and menus; they also take reservations. In addition, the system provides an electronic bulletin board.

The most vibrant infrastructure we can build is one that allows free enterprise to flourish. The telephone companies will play a major role in laying down the physical fibers for the network and so bear the responsibility for providing flexible information transport services, some minimal common services and access to major shared resources. But neither the telephone companies nor any other centralized body should decide what common communications conventions should be offered; that should instead be the role of special-interest groups working outside the purview of the carriers. In addition, no central agency should have anything to say about what services will be offered through the information infrastructure. These new products and services should be devised by the millions of people and their computers that will use the infrastructure according to their own plans and for their own purposes.

The information infrastructure will then resemble an old-fashioned village market. A multitude of goods and services will be bought and sold through this new information marketplace, as I like to call it. As in traditional markets, not all transactions will be monetary. Some people will publish free manuscripts, others will engage in debates and still others will collaborate in creative and entrepreneurial ways, all through the network.

We can begin to imagine the information marketplace by thinking about how it will change familiar information-handling services such as today's business mail. There is no reason to treat much of the mail as precious physical goods. The message is usually more important than the paper on which it is printed. This does not mean we are obliged to send a video clip of shimmering red roses to someone we love. But when the content is more important than the medium, it is wasteful to transport the paper conveying the message.

A speedy and flexible nationwide electronic mail service could dramatically enhance the competitiveness of domestic industries. Whether a nation becomes a major producer of manufactured goods and services depends heavily on the quality and cost of the products it generates and the speed with which the products reach the market. Lowering the cost of production and trimming the time needed to design, produce, sell and service products are all intimately related to the speed and flexibility of communicating and processing business information.

In addition to conventional advertising, manufacturers' computers may respond directly to queries for products with special features and prices broadcast by the computers of potential consumers. Before buying a car, for example, a consumer would interact with the manufacturer's computers: via the infrastructure, she might request certain basic features (namely, four-wheel drive), scrutinize the models that fit, then further tailor the car to her needs and wishes by selecting specific options, fabrics and colors.

Once the sale is made, her order would touch off an explosion of computer network activity. Automatically, her descriptions would expand into a cascade of orders for the necessary systems and subsystems of the car. The order would then generate instructions that would time the arrival of these parts to the production floor and would direct the floor to assemble all these pieces. Such services could lead to the mass production of individualized products. As Thomas W. Malone and John F. Rockart point out, networks may enable independent contractors to create an organization overnight to fill a customer's specific demand—and then to dissolve it just as quickly [see "Computers, Networks and the Corporation," page 140].

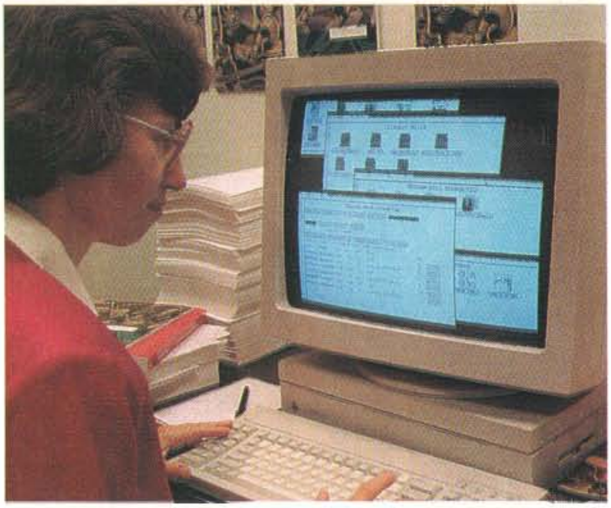
The information marketplace will also change how we work with geographically distant partners. An increasing number of conferences are already conducted over video links, but these conferences still require all the participants to be in the right place at the right time. New approaches to such collaborations will free people to take part in delayed and distributed meetings. Lee Sproull and Sara Kiesler address the social and managerial implications of such "virtual" collaboration [see "Computers, Networks and Work," page 128].

Other services and opportunities also become possible as the infrastructure dissolves geographic and temporal barriers. Remote medical diagnosis and perhaps, some day, remote manipulation would bring experts close to needy patients who are physically thousands of miles away. The long-promised links between the workplace and the home would finally be forged. Besides offering many people the convenience of working at home, the network would also open up new career opportunities for housebound parents of children and for physically handicapped people.

Because the infrastructure so rapidly handles the transport and processing of information, the services likely to gain the most efficiency from the infrastructure are those that involve only information. Securing advice on a specific legal, financial or medical problem could become a relatively easy and low-cost undertaking. Prospective vacationers could explore a virtual island retreat from their home before booking tickets. Inquiries about government regulations would be answered in minutes rather than in months. Advice on building a house or baking a chocolate soufflé would be at hand at any time.

Computers are already heavily used in military training; specially designed flight simulators for pilots and sophisticated war games for strategists have become standard practice. As networked computers grow cheaper and more powerful, companies, too, will be able to afford such approaches. A firm might train its managers by confronting them with simulated problem scenarios involving employees' complaints, project bottlenecks or one another. Other companies might sell such simulation services through the information marketplace.

Even though much has been said and written about computers and education, we have yet to find the most effective ways of applying computers to help people learn. As Alan C. Kay points out, computers do not magically improve schools. When used to drill students on rote memorization, the machines add little value [see "Computers, Networks and Education," page 148]. But by enabling students to interact with rich intellectual resources, including distant teachers, libraries and museums, the infrastructure can be a strong ally in general education.



CUSTOM-TAILORED TEXTBOOKS may help teachers design lessons that meet the specific needs of students. One such effort is McGraw-Hill's Primis software shown here.

In specific cases, we are discovering approaches that seem effective. At M.I.T., students participating in Project Athena learn French as they engage in a simulation of trying to rent an apartment in Paris. A high-resolution moving image of a French speaker asks questions that students must answer.

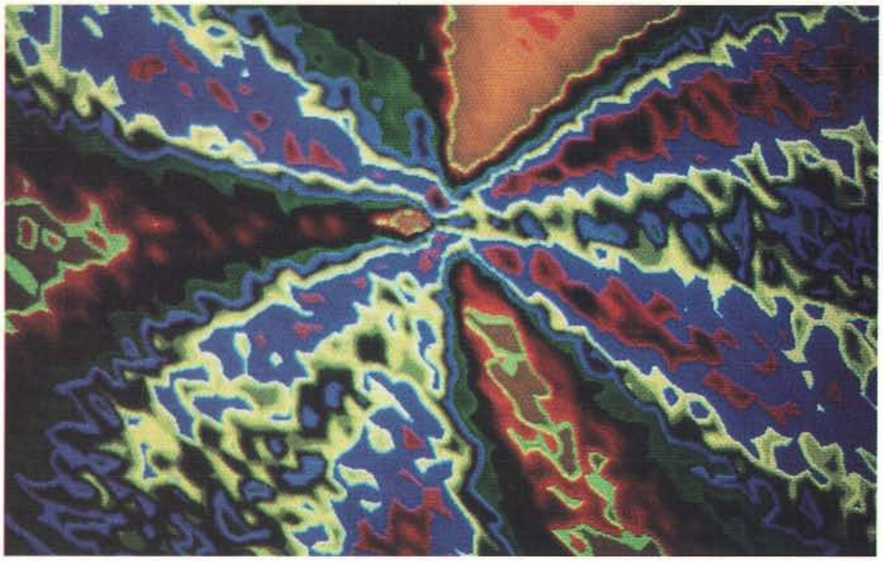
Eventually the infrastructure may even couple individualized computer tutors with students. One such tutor might help a student analyze a specific bridge or, at a more ambitious level, design bridges. If we can make research dreams come true, an even more advanced tutor, endowed with the design approach and style of Frank Lloyd Wright, might help a student become a virtual apprentice of the master.

Entertainment and publishing will be greatly enhanced by the information infrastructure. We will have easy access to millions of movies and music performances by renting them from collectors who offer their goods in the information marketplace. Video broadcasters, radio stations and videotape rental stores may lose some of their monopolistic hue as the information marketplace connects every supplier to every consumer throughout the nation and the world.

Printed newspapers are unlikely to disappear because they are so convenient and inexpensive. But the information marketplace will also be an abundant source of timely information—including text, advertisements, music and video—contributed by anyone and perused freely by anyone else. These mountains of information will, in turn, create opportunities for a breed of electronic publishers and entrepreneurs who will sift for diamonds, which they can then edit and publish electronically. Thus, even though publishers may not print their works on paper, the substance of gathering and reporting information will not change. Nicholas P. Negroponte describes related products and services that will emerge [see "Products and Services for Computer Networks," page 102].

Not only will the information marketplace enhance specific sectors of the economy, but even greater collective benefits will flow from binding together previously unlinked parts of the economy. Consider, for example, the likely evolution of high-definition television (HDTV). With the information infrastructure in place, all the vexing questions about how HDTV will develop and who will control it will fade. Instead we need only adjust the flexible transport levers to make HDTV a reality, linking every home and office to every broadcaster and video rental service in the information marketplace. Interactive, how-to instructions, virtual museum visits or simulation sessions could exploit the visual power of HDTV, thereby benefiting the suppliers of and the many services that use high-definition video as well as the country that has this infrastructure treasure trove.

Where will this information mania lead us? As we think about how far we



"STARBURST #21," a computer-generated work by Lillian Schwartz, represents the artist's view that "what the computer can do is subject to what we believe it can do for us." It appears in *The Computer Artist's Handbook*, published by W. W. Norton.

might use the technologies to stretch the ways people live, learn and work, we anticipate that a sophisticated information infrastructure should help in three ways. It should relieve many of the repetitive, boring and unpleasant tasks related to processing and communicating information. In this case, the computers' effectiveness will be limited only by the extent to which they have been designed to understand the information that reaches them, regardless of whether the information is supplied by people or by computers.

Second, the information infrastructure should help us improve the ways we do things now, by speeding up existing processes or by improving their quality. Although the flow of information can be greatly accelerated, the ultimate limits to these improvements are governed by the physical work involved, including those tasks that only people can carry out. No matter how fast the information flows, assembling a real car out of real parts still takes real time.

The third major way computers and networks will touch our lives is by unlocking as yet unexplored possibilities. The neighborhoods we play in and the people with whom we do business no longer need be the ones close by but the ones we choose.

There is no question, too, that computers and networks will democratize human communications. Nearly everyone would be able to put his or her ideas, concerns and demands before all others. This freedom will undoubtedly bring sociological consequences, including the formation of electronic tribes that can span physical distance.

The manipulation of video, sound and text by computers will let us explore new vistas. It will also further blur the boundary between virtual and real experiences. Even a plain old movie can deeply color the moods of passive viewers. What might tomorrow's interactive, multisensory movies inspire?

Influential as these tools will be, there are obvious limits to how much they will change our lives: no amount of virtual reality can substitute for people's real needs. Neither can computers and networks augment the human capacity to absorb information or the number of people with whom a person can interact or the quality of human relationships.

The information infrastructure will also introduce new challenges for society. Who should be liable for computer services that misbehave? I believe the people or companies who profit by providing the services should be accountable. How will we ensure personal privacy on the network and yet protect users from computer crimes, worms and viruses? We cannot yet answer questions like these with much certainty, but we should not hesitate to try to anticipate them, as both Anne W. Branscomb [see "Common Law for the Electronic Frontier," page 160] and Mitchell Kapor do [see "Civil Liberties in Cyberspace," page 174].

Yet even as some of the promises that we seek will turn out to be mirages, so, too, some of the problems will evaporate like bad dreams. Some people worry that these new technologies will tend to dehumanize us. In the worst technophobic scenario, men and women fixated on computer screens and plugged into networks are rendered spiritless and become trapped in lonely and cruel isolation. I find such concerns implausible. People are neither so naive nor devoid of instincts for self-preservation and control that they will surrender their humanity to their tools.

Information infrastructures will evolve first in those industrial countries that need and can afford them. Vice President Al Gore identifies the role that national governments must take in developing these infrastructures [see "Infrastructure for the Global Village," page 156]. Like the traditional highway and power systems, these information infrastructures will be woven tightly into the fabric of a nation, hard for others to copy or to emulate. And like these earlier infrastructures, they will give their builders certain unique economic advantages.

Once several national information infrastructures are in place, countries will tie them together, much as national power grids, airline routes and telephone circuits have been linked in the past. The result will be a global information infrastructure that will help the people of the world buy and sell information and information services and share knowledge and creative energy—we hope to the benefit of all.

The opportunities along with the problems that may well arise on tomorrow's computers and networks will be new, different, unpredictable and worthy of our continual vigilance. Harnessing the electronic agents that will emerge from this infrastructure to support humanity may be our ultimate challenge. And yet the opportunities and shadows we face as we try to achieve our goals in the information age are consistent with those we have grappled with during the agricultural and industrial eras. The alternative—closing the door to technological discovery so as to avoid societal pitfalls—is unacceptable to the probing nature of the human spirit.

FURTHER READING

THE NATIONAL CHALLENGE IN COMPUTER SCIENCE AND TECHNOLOGY. Computer Science and Telecommunications Board, Washington, D.C., 1988.

SILICON DREAMS: INFORMATION, MAN, AND MACHINE. Robert W. Lucky. St. Martin's Press, 1989.

BUILDING THE INFORMATION MARKETPLACE. Michael L. Dertouzos in *Technology Review*, pages 29–40; January 1991.

M.I.T. PROJECT ATHENA: A MODEL FOR DISTRIBUTED CAMPUS COMPUTING. George A. Champine. Digital Press, 1991.

Domesticating Cyberspace

A technophile and the information, entertainment and communications industries have discovered the source of next-generation electronic products—the network

by Gary Stix

To prepare for a congressional hearing, Vinton G. Cerf sent out an open message on the web of computer networks called the Internet. Cerf, who is president of the Internet Society, asked users how they found the network to be helpful. Within a matter of hours, thousands of replies arrived in his electronic mailbox. "Without the infrastructure of the Internet, asking people about the network would not have been worth it, since the answers would have taken far too long to receive, and it would have been impossible to sort through the masses of paper," Cerf said.

Among the responses was a note from a professor at the University of Southwestern Louisiana recounting how his offer to teach a class using electronic mail relayed over "The Net" elicited responses from some 15,000 prospective students. A blind student told of tracking down archives of William Shakespeare's works that could be read through his speech synthesizer. Another story mentioned how President Bill Clinton's visit to a Silicon Valley computer manufacturer was quickly rendered into a stream of digital bits and transmitted over the network to hundreds of sites.

The creative ferment of the Internet—with its thousands of computer networks, comprising 1.7 million computers in more than 125 countries—is frequently invoked by the legislative legions in Washington who want to extend some version of electronic networking to every home, school, library and hospital in the country under the rubric of a National Information Infrastructure. Not surprisingly, the Internet's 20 percent average monthly traffic increases on its most heavily used segment have also tickled the fancy of media and communications megaliths entranced by an information service with growth rates measured on logarithmic scales.

The explosion of activity on the Internet has coincided with the ascendancy of a vice president who readily poses for photographs beside an ion-beam implanter as well as an unusual confluence of interests among computer, communications and entertainment giants. The migration to the Internet by universities, government agencies, community organizations and even business electronic mail users is seen as stirrings of mass appeal for electronic networking beyond the automated teller machine. "I see seeds of a National Information Infrastructure in Internet," says Robert W. Lucky, vice president of applied research at Bell Communications Research (Bellcore).

Moreover, the technical accoutrements are almost in place to extend the technology of computing into the television set reposing on the living-

GUIDING LIGHT for gigabit networks will be photonic technologies that will become the foundations for a National Information Infrastructure. An AT&T Bell Laboratories researcher monitors the fabrication of a fiber for an optical amplifier that is made by depositing a vapor of erbium into a glass rod.

room credenza—the same trend that saw it move from the mainframe data center during the 1980s to the office worker's desktop. Like the interstate highway system, the main arteries of the "information superhighway" envisioned by Vice President Al Gore have been under construction for the past decade. The flow on AT&T's transcontinental network tops out at 3.4 gigabits per second, or, in the popularized measurement, more than three encyclopedias' worth of text a second. Although data compression technologies can send a digitized movie over existing copper wires, lightwaves may prevail. Corning, the fiber manufacturer, estimates that roughly 2.3 million miles of fiber may be strung in 1993 by leading cable and local telephone companies, a 25 percent rise over last year. (From 12 to in excess of 200 fibers may fit into a single cable, so the actual number of miles of cable laid will be about 55,000.)

What telephone and cable companies propose to build, either together or as competitors, are the "on and off ramps," and the toll booths, to these electronic transcontinental thoroughfares. They will own, if Washington so allows, both the medium and its message units. For the consumer, the cost of admission will be not just the quarter for a pay call but a few dollars more for the video conference, the pizza order and every episode ever recorded of "My Mother, the Car." The $20 or $30 a month for a cable bill could double or triple, once "online connect" charges are added.

To garner the requisite expertise and billions of dollars in financing, the divested Baby Bells and the cable television companies have sheathed their swords to form multimedia *keiretsu* that make an art of the joint venture and strategic partnership. U.S. West teamed up with Tele-Communications, Inc. (TCI), the largest U.S. cable deliverer. Later the same regional telephone company announced an agreement with cable giant Time Warner. Another, Bell Atlantic, has set itself up to be a supplier of video pipes for local cable companies. AT&T is preparing to launch an interactive television trial with the cable company Viacom. And Southwestern Bell has purchased two cable companies near Washington, D.C. "The Mongol hordes are raging, but there is no Genghis Khan nor a David Sarnoff," observes Michael Hawley, a professor of computer science and electrical engineering at the Massachusetts Institute of Technology's Media Laboratory.

Boardwalk and Park Place

Indeed, futurists such as George Gilder of the Discovery Institute, a conservative West Coast think tank, assert that the inexorable pace of technological advance will encourage a wealth of competitors—direct-broadcast satellites, cellular television or hand-held personal communications devices—that will nibble away at the grand old monopolies held by the offspring of Ma Bell. In Gilder's view, the existing regulatory framework is antiquated and should be scrapped so that companies can form partnerships or engage in a free-for-all as they choose. But that has not yet happened. The cable industry has just been reregulated because of service problems, rate gouging and the failure of competition to emerge in local markets. And the basic connections to the home on the telephone company's local network are still a protected preserve.

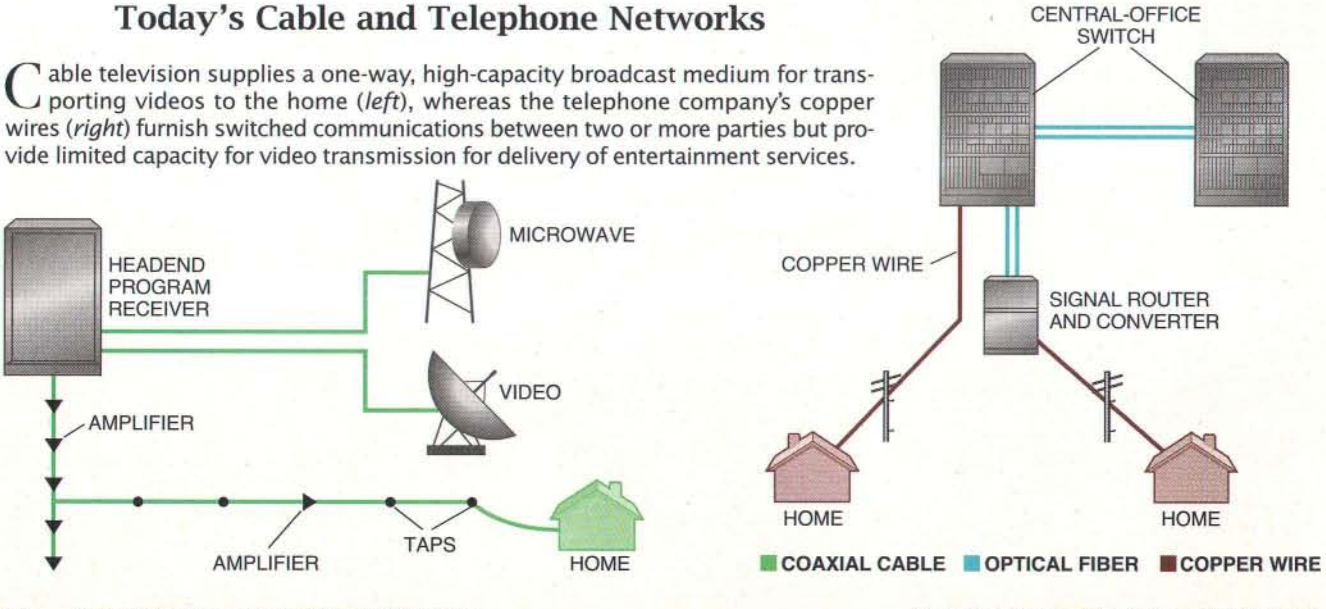
Even before Gore put forward the notion of an electronic superhighway, cable companies were beginning to use optical fiber to improve their existing networks. Their goal was to eliminate the string of failure-prone electronic amplifiers that are needed to boost the signal every 2,000 feet—one reason that cable systems often have outages. "In a large, urban cable television network, the distance to the furthest subscriber is about 40 amplifiers," says John Ryan, a South San Francisco communications consultant. "That's a maintenance nightmare." Just as important is the fact that the amplifiers induce noise into the network, which affects picture quality. This equipment also inhibits the two-way communication needed for ordering videos or products hawked on a home-shopping show.

In the upgraded networks, optical cables reach a point a mile or less from a group of homes. Then existing coaxial cable provides the rest of the link. In raw communications capacity, this cable supplies the industry with an advantage over the narrow bandwidth in the telephone network's copper wiring. The coaxial cable can carry up to 1,500 channels of compressed video signals for short distances. The cables that constitute the physical pathways of this terrestrial television network already pass about 95 percent of U.S. homes and serve almost 60 percent of all households, according to market research estimates.

After the election of Clinton and Gore, some of these capital upgrade programs were politically corrected to become information superhighways. The "Infostructure Network" is the trademark that TCI has applied for. The company announced in April a four-year, $2-billion capital-spending plan to lay fiber

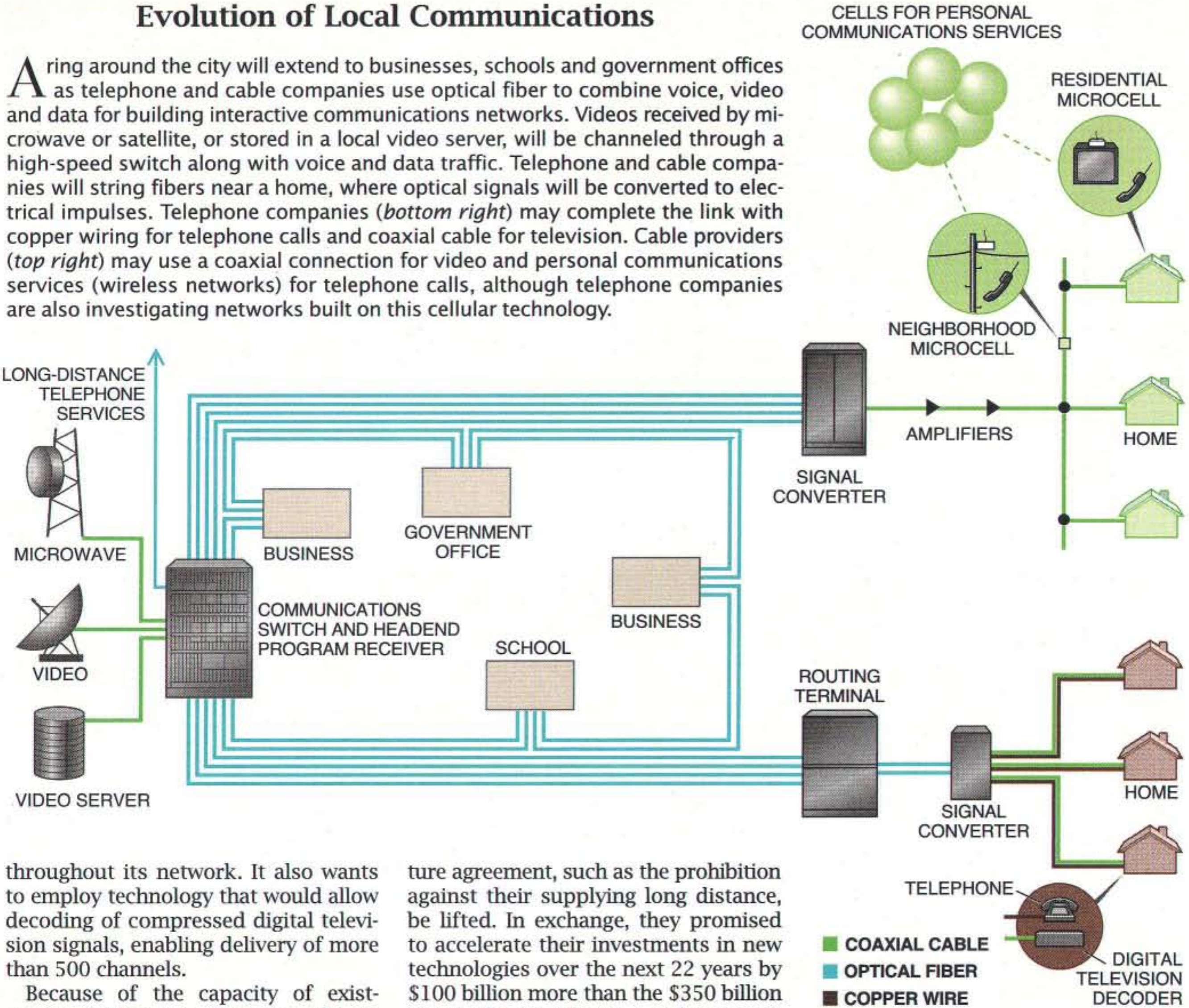
Today's Cable and Telephone Networks

Cable television supplies a one-way, high-capacity broadcast medium for transporting videos to the home (*left*), whereas the telephone company's copper wires (*right*) furnish switched communications between two or more parties but provide limited capacity for video transmission for delivery of entertainment services.



Evolution of Local Communications

A ring around the city will extend to businesses, schools and government offices as telephone and cable companies use optical fiber to combine voice, video and data for building interactive communications networks. Videos received by microwave or satellite, or stored in a local video server, will be channeled through a high-speed switch along with voice and data traffic. Telephone and cable companies will string fibers near a home, where optical signals will be converted to electrical impulses. Telephone companies (*bottom right*) may complete the link with copper wiring for telephone calls and coaxial cable for television. Cable providers (*top right*) may use a coaxial connection for video and personal communications services (wireless networks) for telephone calls, although telephone companies are also investigating networks built on this cellular technology.



throughout its network. It also wants to employ technology that would allow decoding of compressed digital television signals, enabling delivery of more than 500 channels.

Because of the capacity of existing coaxial links to the home, the cable industry may be able to rebuild its network with fiber for $200 to $300 per customer, whereas telephone companies may need to spend $1,500, according to market and academic researchers. That differential may narrow, though, when cable companies must go beyond merely broadcasting their programming, as they do now, to furnishing switching services that can route a voice call or a video conference from one point on their network to another, as does the telephone company. Cable companies may also have to learn something about customer service. They have not been compelled to maintain the same level of service reliability as the telephone companies. Having access to 911 is deemed more important than being able to view "The Simpsons."

Meanwhile the regional telephone companies have been spending a lot of time in Washington since the election. Top executives from the Baby Bells met with Gore in mid-April to ask that remaining restrictions of the Bell divestiture agreement, such as the prohibition against their supplying long distance, be lifted. In exchange, they promised to accelerate their investments in new technologies over the next 22 years by $100 billion more than the $350 billion they had originally expected to spend. Reaction to such proposals has been mixed. "If you let local telephone companies into long distance, that's all you're doing—putting Humpty-Dumpty back together again," MCI Communications chairman Bert C. Roberts, Jr., said in a speech.

So far the Clinton administration has given no official response, but the progeny of Ma Bell have already gotten much of what they wanted. The Federal Communications Commission opened the door last year for telephone companies to carry video programming over their networks. But a prohibition against owning or licensing programs in the companies' own serving areas stands. After the FCC decision, New Jersey Bell, which has plans to pave virtually all of its network with fiber by the year 2010, struck an agreement with two cable companies to provide this "video dial tone."

Despite wider regulatory latitude, the Baby Bells are not satisfied. Bell Atlantic, the parent of New Jersey Bell, wants to become a cable company and has set up a separate subsidiary called Bell Atlantic Video Services. Along with Chesapeake & Potomac Telephone Company of Virginia, another operating unit, it sued the federal government last December, charging that the programming prohibition violates the company's First Amendment rights. The newspaper, broadcast and cable industries have filed friend-of-court briefs in opposition. Two bills in Congress would also release the telephone companies from these programming restrictions.

At the same time, the telephone companies' technical ability to compete with the cable networks is slowly improving. In many cases, they already have fiber pipes connecting their switching offices. Fiber constitutes almost 5 percent of total cabling (the rest is copper) in the Baby Bells' network. In extending these links, fibers will not stretch all the way to a home but to a curbside attachment point, or optical network unit, a few

The U.S. Lead in the Information Age

TELEPHONES	U.S.	JAPAN	EUROPE
LINES PER 100 PEOPLE	48.9	42.2	42.2
CALLS PER PERSON PER MONTH	43.4	46.1	48.7
CELLULAR TELEPHONES PER 100 PEOPLE	2.6	1.2	1.2
TELEVISION			
HOUSEHOLDS WITH CABLE (PERCENT)	55.4	13.3	14.5
VCR-RELATED EXPENDITURES PER HOUSEHOLD PER YEAR IN DOLLARS	44.6	35.3	14.1
COMPUTERS			
PERSONAL COMPUTERS PER 100 PEOPLE	28.1	7.8	9.6
DATA-BASE PRODUCTION (PERCENT OF WORLD)	56.0	2.0	32.0

SOURCE: Consumer Federation of America

hundred feet away. From there, coaxial cable will bring the video signals into individual homes; copper wires will carry in the telephone signals. A North Carolina company, BroadBand Technologies, has introduced the electronics for a fiber transmission and switching system that forgoes the complex modulation and circuitry used to broadcast cable programming. Similar to the technology used in some computer local-area networks for offices, it may be able to offer interactive video to the home for less than $500, the company claims.

Yet the telephone companies have still to deliver on the promise of some investments already made. A network that could seamlessly switch digital voice, data and video—the essence of an electronic superhighway—has been talked up for a decade by the Baby Bells as part of a scheme known as Integrated Services Digital Network. ISDN for the home would supply two digital channels, each 64,000 bits per second (the amount of bandwidth required for an uncompressed digital voice signal), and an additional data channel of 16,000 bits per second. Besides giving the consumer two telephone lines, these digital channels would dwarf the capacity of the analog modems that now transmit digital signals along telephone lines at speeds up to 14,400 bits per second.

The digital switches and software for ISDN are in place in many telephone company switching offices, paid for out of basic telephone rates. Although ISDN may furnish most of the capabilities of a broadband-fiber network, the telephone companies have lagged in deploying the technology at affordable rates. Today five states have ISDN tariffs targeted to residential customers, and costs can be high. Telephone companies say the service will become more widely available now that a set of technical standards is in place that lets switching equipment from one manufacturer work with that of another in routing ISDN's digital signals.

Mark N. Cooper, director of research for the Consumer Federation of America, argues that the telephone company's focus on optical-fiber technology will cost billions and lead to overcapacity for already underutilized digital switches. Nor is it necessary to build these broadband networks to keep the U.S. ahead of foreign competitors. According to the federation, the U.S. has the world's most advanced telecommunications infrastructure, outpacing Europe and Japan in the penetration of telephone lines, cellular telephones, cable television and personal computers. Especially significant is the fact that the U.S. population has absorbed many times the number of personal computers as have the Europeans or the Japanese. "The argument that pulling fiber will advance them ahead of us when nobody there has PCs is a non sequitur," Cooper says.

Low-Fiber Diets

The sense of urgency to deploy fiber may be further mitigated by asymmetric digital subscriber line (ADSL), a technique developed by Bellcore, the research arm of the local telephone companies. ADSL uses specially designed transmitters and receivers to reduce the interference and noise normally encountered on a copper line. It would permit transmission of compressed video signals over ordinary copper wire—at best, superior in quality to a conventional television picture. The ADSL transceiver can be attached to a single subscriber line without having to string optical fibers to an entire neighborhood.

The Electronic Frontier Foundation (EFF), an organization established by software entrepreneur Mitch Kapor, has become an advocate for these digital technologies because they allow a mail message or a digital video file to be routed from any sender to any receiver with the telephone network's switching capabilities—an inherently democratic communications medium. "It makes people into information publishers as well as consumers," says Daniel J. Weitzner, the EFF's senior staff counsel.

Such a switched architecture contrasts with the cable network that broadcasts one program to many receivers. Vestiges of this hierarchical system may persist even after cable networks permit two-way communication over their television networks or after the telephone companies build optical links, Weitzner says. "There's a concentration of enormous economic power behind this media distribution," he notes. "What we may be looking at is one-way entertainment coming down through the pipe with a little upstream capacity to select a movie or an item out of the Victoria's Secret catalogue."

Some networking experts believe that anything but broadband pipes will quickly be exhausted. "You only get a chance to build new infrastructure every once in a while," says Rick Weingarten, executive director of the Computing Research Association, which represents academic and corporate researchers. But one key legislator who agrees with the EFF's position is Representative Edward J. Markey of Massachusetts, chairman of the House Subcommittee on Telecommunications and Finance. He has stated that he would like to see digital services widely deployed before his retirement in 2015, the date that telephone companies would finish laying fiber throughout the local telephone networks under an accelerated schedule. "The goal is not fiber to the home but information to the home," Markey says. Before the year's end he hopes to introduce legislation that will require the regional telephone companies to file tariffs at affordable rates for digital services, such as ISDN.

Whether or not the National Information Infrastructure ends up being born optical or relies on digital technologies that can utilize existing copper wires and coaxial cable begs the question of who will pay—and how. And here, too, the ad hoc Internet has been drawn into the center of the debate over the appropriate role that government and industry should play in building and operating communications networks.

This past March executives from the major local and long-distance telecommunications companies signed a policy statement that outlined their vision of the place for government and the private sector in constructing networks. Arguing that government subsidies for building networks will deter commercial providers from entering these markets, the telecommunications industry broke down networking into two categories: "experimental" and "production."

Experimental gigabit networks could receive direct government funding. Production networks, those tapped into by researchers who do not require such advanced communications capabilities, would be built and operated entirely by the private sector. Federal funding for building the academic and research networks used for routine communications such as electronic mail would be phased out, although the government could subsidize users who purchase connections to the network from commercial providers.

Some of these suggestions are in line with what the government was already planning. In coming years, the National Science Foundation (NSF) wants traffic traveling over its "backbone" network—today the main conduit for the Internet—to be carried by commercial suppliers. Instead of the existing backbone, the agency has commissioned a high-speed network that will be limited largely to research on communications and supercomputers.

The telephone industry's well-defined categorizations have riled leaders of organizations that establish and manage the various "mid-level" networks run by universities, research institutes and libraries. The NSF, though, will continue for a time to provide a needed measure of funding for building and operating some of these networks. This dispute, it is feared, could ultimately disrupt the spontaneous, grass-roots quality that has enabled the Internet to grow into a worldwide network serving some 10 million users.

The telephone industry has been lobbying hard in Congress to make sure the public-private distinctions are retained in various bills to fund applications for health, education and library services under the rubric of the National Research and Education Network (NREN). If the Internet—or some outgrowth of it—evolves into a network that connects schools, hospitals and other public institutions, the telephone companies believe they may have a role in providing the kind of services they fur-

Stories from "The Net"

Earlier this year the Federation of American Research Networks, a nonprofit group that promotes use of computer networks in research and education, asked how the Internet can be of value to schools, health care and various public institutions. Here's a sampling of what people said.

Bruce Daley, teacher, Elaine Wynn Elementary School, Las Vegas, Nev.: "Paul Smith, a 24-year-old doctoral student from Latrobe University in Melbourne, Australia, was recently assigned to Casey Research Station on the continent of Antarctica. While stationed at the research facility, Smith used electronic mail over the Internet to describe his experiences of living and working in Antarctica to my third-grade class."

Norman F. Herzog, manager, engineering department, UNO Television, University of Nebraska at Omaha: "My daughter was scheduled for surgery in October of 1991 for correction of scoliosis (curvature of the spine). In late summer of that year, I decided it was important to learn more about scoliosis. A library catalogue search over the Internet led me to discover that another daughter had symptoms that could mean our family was affected by a serious hereditary disorder known as Marfan syndrome. I used a specialized Internet service, called WAIS, that let me search multiple data bases. The bibliographies led me to physicians who knew how to diagnose and treat it. The Internet may have saved my daughter's life!"

Allan G. Farman, professor, radiology and imaging sciences division, University of Louisville: "In dentistry, approximately 20 percent of all procedures require prior approval based upon the submission of radiographs. The Internet has proven very efficient and reliable for transmission of the relatively small amount of data in dental images. Transmission rates of 20 seconds have been possible for images between Louisville and Seattle, Vancouver, Torino, Italy, and Amsterdam, the Netherlands. Image analysis has shown that there is no loss of information during transmission."

Ferdi Serim, computer teacher, West Windsor/Plainsboro Regional Schools, Plainsboro, N.J.: "I strive to make the classroom a place where children feel comfortable exploring, and through the Internet our 'playground' is truly global. One example was the Chernobyl exercise. It lasted only 36 hours, but it was profound. On Monday, at noon, a message came over the Internet stating that 40 students from the contaminated zone in Chernobyl would be visiting a health spa in England. On Tuesday morning, my first two classes used our word-processing skills to craft 'get well' messages for the children. It took a little more than an hour to write 50 messages and 30 seconds to send them to England. The next morning we received a reply from the Russian teachers. They were amazed that U.S. children knew and cared about the plight of their students. History, science and language arts blended to make a personal sense of current events tangible to our students in a way otherwise impossible."

Ferdi Serim's computer class

nish on their large voice networks: reliable access and sophisticated billing capabilities. To make their point, they have named managers with titles like marketing director for Internet/NREN and executive director for National Information Infrastructure Initiatives. "This is a delicate thing," says Bellcore's Lucky. "The Internet is a wonderful sociological experiment. The question is, can it be scaled up to an industrial-strength network and still retain its flavor?"

Internet veterans point out that telephone companies have yet to deliver. The Baby Bells generally do not supply the types of expertise and services—software to route messages on the network, for example—that users expect from smaller networking organizations. "These markets may not have a lot of money in them," says Weingarten of the Computing Research Association. "They're not well organized, and they are not attractive for the telcos to go hitting on." Others guess that the communications services market may, in fact, exceed $100 million nationwide, but the telephone companies' staidness is keeping them away. "They don't have the trained staff—they don't have the experience in operating these kinds of networks," Cerf charges. "What they're offering is a pack of excuses instead of getting in and getting their feet wet."

While sparring continues over the Internet's future, the media-marketing empires are forging billion-dollar deals. Their reason is not solely an altruistic urge to better serve needy schools and hospitals but the belief that traditional broadcast and print media will merge into a multimedia cloud that can float into the consumer's head during the more than seven hours a day the television is switched on. "The country needs an information supermarket, not an information superhighway," says Arno A. Penzias, who is AT&T Bell Laboratories's vice president for research.

Yet no one knows whether consumers are truly interested in foraging through a video wilderness of 500 to 1,000 channels. The idea that people want a virtually infinite choice of programming, not 10 or so channels, strikes some traditional broadcasters as an absurdity. "They won't buy the accountancy channel, the plumbing channel or the hardy perennials channel," Howard Stringer, president of CBS/Broadcast Group, told an audience at a broadcasters' convention this past April.

Reasonable people will differ. "There's a magazine called *Guns and Ammo,*" Penzias says. "There are people who will sit there all day reading about a bullet. As far as they're concerned, if there is a channel devoted to guns, their programming will improve."

This kind of frenzied activity has a cyclical quality. In the late 1970s the Knight-Ridder company assembled in great secrecy a team of editors and technologists to create Viewtron, a service with community news, banking services, travel reservations—many of the things that interactive television promises, with the addition of the video. A fatal mistake then was giving customers a terminal from AT&T that was incompatible with the Apple and IBM computers that were gaining widespread acceptance.

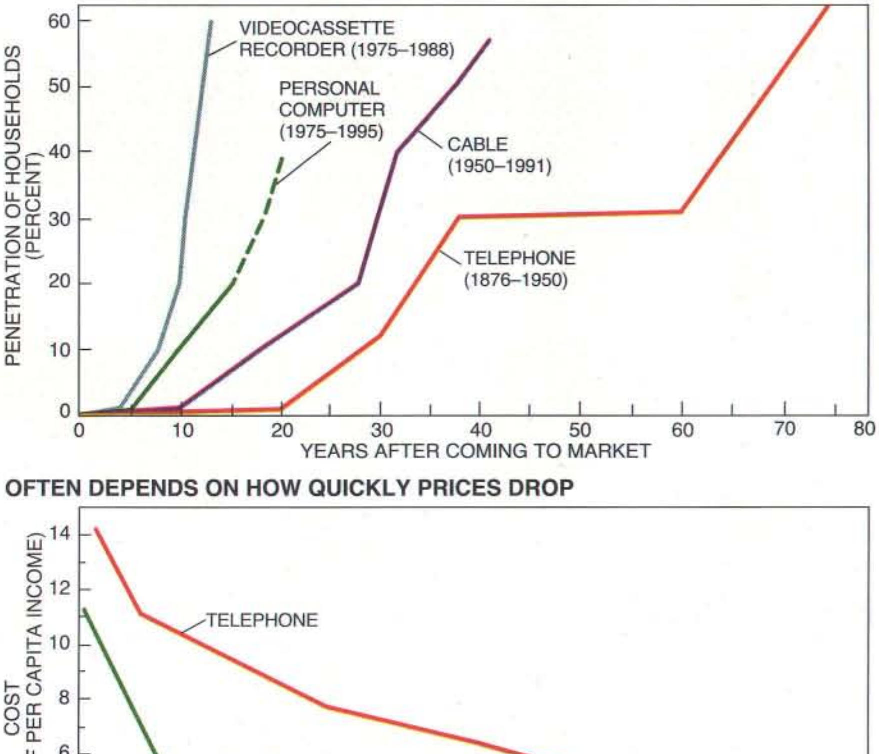
In 1986, after spending $50 million, Knight-Ridder dropped its experiment. Viewtron subscribers primarily logged onto the system for electronic mail and bulletin boards that now draw personal computer aficionados to on-line services like CompuServe and Prodigy as well as to the Internet.

Déjà Viewtron

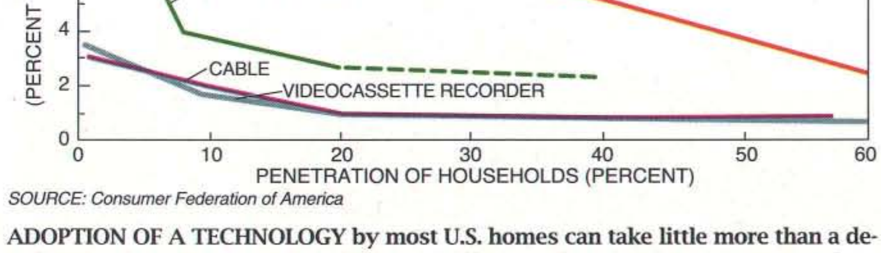
But spurred by improvements in the cost and performance of computing technology, Knight-Ridder has come back for more, although it is treading cautiously. For some time, it has supplied on-line text versions of its newspapers. Last fall it set up an information design laboratory in Boulder, Colo., that is exploring the concept of an all-electronic newspaper that goes beyond sending plain text files by modem. This time the group has no plans to commission special hardware, as it did before from AT&T. But it is experimenting with early versions of a two-pound, tabletlike computer consisting of a flat-panel screen about the size of a standard sheet of typing paper.

A penlike pointing device lets a reader unpack a full news story from a page full of abstracts, similar to the *Wall Street Journal*'s front-page news summaries. These stories could be beamed to an electronic tablet over a radio link. The reader might then access a number of other stories, old or new, in the newspaper's data base: sidebars on dif-

HOW FAST THE INFORMATION AGE COMES HOME...

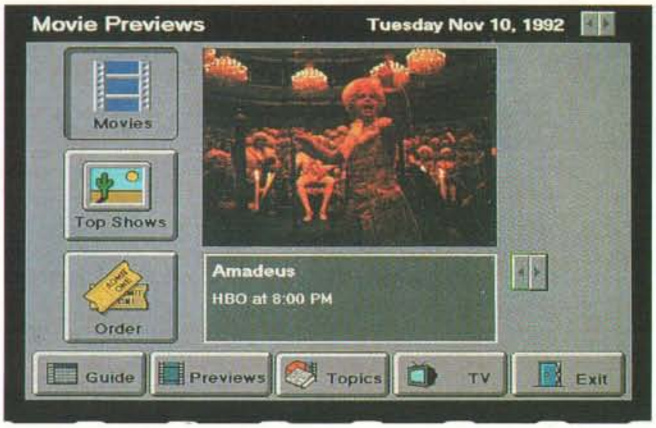


OFTEN DEPENDS ON HOW QUICKLY PRICES DROP



SOURCE: Consumer Federation of America

ADOPTION OF A TECHNOLOGY by most U.S. homes can take little more than a decade or as long as 75 years (*top chart*). Diffusion often accelerates as costs drop (*bottom chart*)—a reason interactive services may now prove more successful.



Amadeus
HBO at 8:00 PM

TALKING BACK TO THE TV may become the great American pastime if electronic video stores and newspapers take hold. Roger F. Fidler of Knight-Ridder's Information Design Laboratory is leading an effort to create a prototype electronic newspaper using a computerized tablet (*left*), and Microsoft has developed a working model of an interactive movie guide that would run on a cable television version of its popular Windows operating system (*right*).

ferent ethnic groups in Bosnia, maps of the region and short video clips.

The secrecy that surrounded Viewtron is gone. Today there is a premium on risk sharing. Knight-Ridder teamed up this past spring with several other media companies—Gannett, Hearst and Times Mirror, among them—to support the "News in the Future" program begun by the M.I.T. Media Laboratory with $6.5 million over five years. "We're trying to make news understandable to the machine," says Nicholas P. Negroponte, the Media Lab's director.

But the integration of artificial-intelligence techniques into the process of editing and reading a newspaper is poorly understood and controversial. The Media Lab has been a leader in the development of "agents," a kind of software version of the television character Max Headroom. The agent, in principle, will watch a reader, note his or her preferences and then go out and dredge up news items from the terabytes of information in data banks, the makings of a personalized newspaper. "The idea of an intelligent agent asking whether you would like french fries with that burger is something that people don't really like," says Diana Hawkins, a former Media Lab researcher, who heads her own firm, Interactive Associates.

To Roger F. Fidler, a member of the original Viewtron team, agents may yield an improved method for reporters and editors to ferret out information from large stores of information. The actual task of editing a newspaper—the decisions by the editor about what news is important—is a form of nonsequential organization that surpasses what may be accomplished by an agent or other computer techniques to link data. "One of the things not well understood by technologists is that editors spend a good deal of time anticipating what readers want," Fidler says. "People think you can just hit this infinite 'more' button, and things will just materialize out of cyberspace."

Overlaying proved and familiar elements—lead story, headline and lead paragraph—onto the amorphous structure of the electronic medium may turn out to be the best way to wean readers away from cellulose, Fidler believes. The saving grace of the electronic newspaper may be institutional survival.

During the 1960s and 1970s, newspapers decided to invest in computer technology to reduce the high costs that were making them uncompetitive against radio and television. "Today newspapers up to the printing press are digital," Fidler says. "The vestige of the industrial age is the printing press, and that is where the next opportunity lies to make enormous savings."

Despite the rush toward digital media, Penzias's vision of an information supermarket may be years away. And it may bear little resemblance to an at-home video store or an electronic news broadsheet. But what has begun to emerge are the bits and pieces of tools for handling any type of digital information, particularly the intensive processing and storage needs of video and film. These trusses and pillars of the information infrastructure could help spur a reemergence of a U.S. consumer electronics industry, since most of the technology is based on the innards of personal computers and workstations.

Moreover, the information revolution will almost certainly be televised. Simply concentrating on building TV sets that produce a crisp, photograph-quality image, once thought to be the cornerstone of the next generation in consumer electronics, is already beginning to take a backseat to incorporating computing power into home video displays. A laggard in getting involved with high-definition television (HDTV), the U.S. came up with digital technologies that ultimately won out over analog transmission systems being considered as standards in Europe and Japan. In 1990 the General Instrument Corporation devised an ingenious HDTV prototype that combined a variety of compression technologies to cram a digital video signal into the six megahertz of bandwidth allocated to broadcasters.

High-Definition Computing

Other companies and institutes subsequently turned their efforts to developing digital systems. At the same time, the computer industry argued for standards that would make the new sets compatible, in all respects, with the computer. This should encompass the way the cathode-ray gun paints the lines on the screen, the shape of the picture elements and the encapsulation of data into small packets. "The next generation of technology should take into account up front that the television is a digital display device," says Lee McKnight, who directs M.I.T.'s program on digital open high-resolution systems.

At the National Association of Broadcasters meeting in mid-April, Apple chief executive John Sculley suggested that HDTV's bandwidth could be broken up into smaller channels as needed, an "intelligent data stream" for programs with multiple camera views or text inserts. A viewer might look at a newscast on the environmental impact of a new dam by clicking a remote control button to choose the viewpoint of a construction worker or that of a fisherman. "I think that may be at least as interesting if not a more interesting opportunity during the 1990s than HDTV itself," Sculley noted.

By late May the computer companies seemed to have gotten at least some of what they sought. The major groups developing systems for testing under the FCC's aegis announced they had agreed to put forward a single digital standard for HDTV. Their proposal is a compromise that accommodates both interlaced scanning—one in which every other row is drawn on the screen by an electron beam—and the progressive scan (in which each row is successively traced), the format desired for compatibility with computers.

The interlaced display causes enough flicker to produce eyestrain for someone looking at text or graphics on a television screen. "Ultimately, your entertainment screen is going to have to be a display for all sorts of data access," Penzias declares. "Under those circumstances when you go to do some data processing, basically the display has got to be able to support a video that may be a dirty movie or data visualization, but it's got to be able to do both."

To do that, others are figuring out ways to imbue the household television with the ability to process the protoplasm of digital bits. What Gilder of the Discovery Institute calls a "telecomputer" is already provoking a competitive struggle. As with the personal computer, which undercut the primacy of central mainframe processing, the design and the amount of intelligence in the box on top of the television will determine how much control the consumer has over interactive television.

Dominant hardware, software and service suppliers want to make sure that their products become market standards. Microsoft, Intel and General Instrument are incorporating the equivalent of an IBM PC running a special version of the Microsoft Windows operating system into a cable converter box. In June, Microsoft was negotiating with TCI and Time Warner about joint development of an operating system for interactive television. Another company, Silicon Graphics, plans to make a television set-top box with the fast-processing chip used in its workstations.

The myriad deals being struck on all fronts are already roiling the regulatory waters. In March, Ameritech, a regional telephone company, petitioned the FCC to begin a process that could result in the lifting of legal prohibitions against offering long-distance service. In exchange, Ameritech would open that bastion of monopoly, the local telephone network in the five states it serves, to competition. Each part of the telephone network could be purchased separately: a customer could order a dial tone from another provider while still leasing the line itself from Ameritech and retaining the same telephone number. Columbia University telecommunications economist Eli M. Noam has suggested that "personal systems integrators" might be needed to make sense of the various options for consumers.

Ameritech and other telephone companies argue that they have nothing to lose: their businesses are under siege. There is already competition for yellow pages, operator and 800 service, pay telephones and regional toll calling. So-called competitive access providers can connect a large business directly to a long-distance carrier at an often lower rate than can the local telephone company. In Houston, for example, U.S. West could be competing against Southwestern Bell in helping Time Warner convert its cable networks to two-way communications systems.

This technological assault, predict analysts such as Noam, erodes the tradition of the common carrier—the obligation to maintain a public viaduct for a free flow of information, open to all comers. Competitors to monopoly common carriers such as the telephone companies can offer a lower pricing structure because they do not have to serve all customers on an equal basis. In tandem with the notion of common carriage, federal and state regulators have required that telephone companies furnish service in high-cost rural areas and offer affordable rates to low-income customers. In turn, urban and middle- to upper-income customers have paid more to ensure universal service.

Civil libertarians worry that having a common carrier like the telephone company own both the medium and the content—which, except for video programming, is now permitted—may stifle both competition and free speech. "Providers of a conduit that have an economic interest in the content running through that conduit are likely to favor themselves in unfair and anticompetitive ways," says Nicholas Johnson, an FCC commissioner during the Johnson administration. "It doesn't matter whether it's coal on a railroad, petroleum running through a pipeline or cable television on a telephone wire."

Consumer advocates argue that competition is more a specter than a reality: competitive access providers that connect businesses to long-distance carriers account for far less than half of 1 percent of the $95-billion local telephone market. Safeguards, they assert, are needed against monopoly predations. "It's like saying that technology will win over everything, so we should just back away," says Gene Kimmelman, legislative director of the Consumer Federation of America. "In 1984 we were told that cable should be deregulated because there were satellites, and all kinds of technology was coming, so we pushed regulation out of the way, and cable became an entrenched monopoly."

Congress, regulators and the courts have yet to grapple with the rewriting of communications laws needed to bring common carriage and universal service obligations to cable companies and other prospective suppliers of two-way communications. A multimedia information superhighway also suggests the need for a substantive rewriting of public broadcasting law. "We're really afraid that all we're going to be left with is a virtual shopping mall," says Jeff Chester of the Center for Media Education, which advocates the taxing of the various new media to pay for noncommercial programming.

Information for All

The idea of universal service may eventually encompass information. Access to the network may be required—not just for calling the fire department or for ordering pizza but for voting and participation in town meetings. "Information stamps" might be parceled out like coupons for buying food. "The equity issue is a very important one in this whole scheme," says Charles M. Firestone, director of the Communications and Society Program for the Aspen Institute. "These technologies could bridge the gap between rich and poor, and they could also, I'm afraid, widen that gap."

The tug and pull of competing interests may mean that little gets done on Capitol Hill or in the state utility commissions during the current sessions. Congress is likely to pass legislation to permit the auctioning off of 200 megahertz of radio spectrum to the private sector for wireless technologies such as personal communications services. Bills for networking for schools, libraries and health care may get through. The Clinton administration was also preparing to appoint a cross-agency task force to deal with computing and networking.

But a definitive road map for a National Information Infrastructure is still lacking. In its absence, the vitality of a public medium like the Internet could be jeopardized as powerful media interests jockey for position. "Anything is a danger in cyberspace," Representative Markey comments. "There are no rules. It's the Wild West." And claim stakes on the electronic frontier may be sold in 30- and 60-second spots.

“The pen isn’t the only thing mightier than the sword.”

Bruce Cunningham
Lanman Companies
Washington, D.C.

Founded in 1911, Lanman Companies is one of the largest privately held digital imaging and prepress operations in North America, with clients like the American Automobile Association, General Mills, Harcourt Brace and Company, the National Geographic Society and *Smithsonian* magazine.

"Change seems to be the big 'new idea' these days. But our business has been changing for years. That's why we're using Power Macintosh to embrace new technologies like multimedia–so we can survive, and thrive. It also helps us integrate our services, improve efficiency and slash budgets. You know, maybe Power Mac could change the way they do business on the Hill."

Using Power Macintosh™ and databases full of scanned and digitized images like this Lenox china clock, Lanman is well positioned to help clients with digital catalogs, interactive kiosks, ads on the Internet and more.

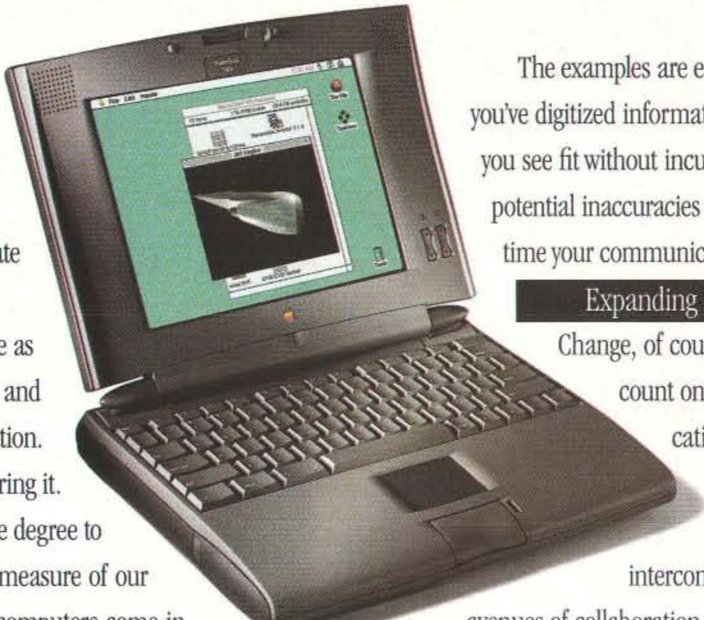
Power Macintosh. It runs optimized applications at blistering speed. It runs Windows and DOS. It boosts productivity. And lowers costs. For companies like Lanman. And for companies like yours. For more information or the name of your nearest authorized Apple reseller, call 800-554-3848, ext. 150, today.

"To test productivity, we didn't just clock application speed. We took a typical catalog spread, accounted for disk access, bus time, overhead, all that stuff, and still ended up with a 26% gain–more than a day a week. Which means this: buy Power Mac. Or lose money."

Power Macintosh™ The business Macintosh™

Apple

The power of information lies in using it.



In this so-called information age—an age of information superhighways, information on demand and information overload—the last thing anyone has time to do is re-create information that already exists.

After all, we spend enough time as businesspeople, scientists, educators and consumers simply gathering information. Dissecting it. Communicating it. Sharing it. Illuminating it. Reshaping it. And the degree to which we do all this effectively is a measure of our success. Which is where Macintosh® computers come in.

Making better choices.

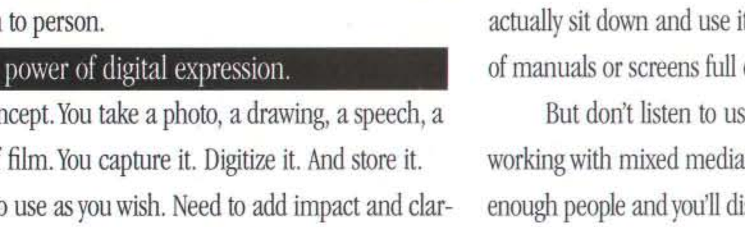
With a Macintosh you can communicate in writing. You can say a thousand words or so with a single illustration. You can create slide shows. Sound bites. And movies. You can present information on paper. On film. On a kiosk. On the Internet. Or right on your own PowerBook® screen. The choice is yours.

But that's nothing new, is it? What *is* new is how the powerful combination of Power Macintosh and software innovations like QuickTime® and the Apple® Media Tool makes it so easy to "re-purpose" information on the spot. So you can always present information in the most appropriate way, even when that way changes from moment to moment or person to person.

The power of digital expression.

It's a simple concept. You take a photo, a drawing, a speech, a recording or a clip of film. You capture it. Digitize it. And store it.

Now it's yours to use as you wish. Need to add impact and clarity to your next technical symposium? Simply collect all the images, sounds, text and film clips you need, and build a multimedia show. Need a detailed printed report? Just use the same data. Need to create a looped VHS videotape? Just gather the goods and print directly to tape.



The examples are endless, the idea simple: once you've digitized information, you can re-purpose it as you see fit without incurring the inefficiencies and potential inaccuracies of starting from scratch every time your communications needs change.

Expanding spheres of influence.

Change, of course, is the one thing you can count on when it comes to communications. With the explosion of online services such as Apple's eWorld™ and their growing interconnection via the Internet, new avenues of collaboration are arising. And the savvy communicator must learn quickly how to travel them. As technologies like videoconferencing gain use, the need to excel at mixed-media communications becomes critical. And, finally, as your competitors become more adept at utilizing new media, so must you.

The power of Power Macintosh.

It will come as little surprise to you that we believe Power Macintosh is the computer of choice in this mixed-media environment. Its RISC-based PowerPC™ microprocessor, for one thing, provides the raw horsepower you need to manipulate complex graphics, digitized video and sound. And because it's a Macintosh, you can actually sit down and use it without slogging through massive stacks of manuals or screens full of error messages.

But don't listen to us. Ask any creative professional currently working with mixed media which computer he or she prefers. Ask enough people and you'll discover what we have: that 63% of all multimedia titles (even those designed to run under DOS and Windows) are developed using Macintosh.

For more information by fax about Power Macintosh and other powerful multimedia tools from Apple, pick up the phone and call us today at 800-770-4852. Or, if you prefer, contact us via the Internet at http://www.info.apple.com.

Apple

©1995 Apple Computer, Inc. All rights reserved. Apple, the Apple logo, Macintosh, PowerBook and QuickTime are registered trademarks of Apple Computer, Inc. eWorld, the eWorld logo, Mac, Power Mac and Power Macintosh are trademarks of Apple Computer, Inc. PowerPC is a trademark of International Business Machines Corporation, used under license therefrom. Windows is a trademark of Microsoft Corporation.

A Digital Fix for the Third World?

Information technology may bring limited benefits to the work force of developing nations

The notion that technology can fix the economic problems of the developing world dates back to the early years of the post-colonial era. The "green revolution" of the 1960s held out the hope that the labors of Western plant geneticists could eliminate food shortages by raising crop yields.

The information age has begun to raise similar expectations. African heads of state, such as Flight Lieutenant Jerry J. Rawlings, who staged the overthrow of a civilian government in Ghana more than a decade ago and was elected as president in 1992, sound a little like executives from Silicon Valley. Rawlings now talks about an "enabling environment" for efficient telecommunications in a country where many employees spend about half their time going from one place to another because of the unreliability of the telephones.

Rawlings's neighbor, Felix Houphouet-Boigny, longtime leader of Ivory Coast, voices a similar hopeful refrain: "Thanks to informatics, technological shortcuts to development exist today and are within the reach of everyone. We are not therefore doomed to remain undeveloped forever."

For all its promise, the allure of skipping over several generations of technology has fallen into disfavor in some circles. Economists often adhere to the "product cycle" approach to diffusion of technology. This trickle-down theory holds that developing nations absorb new technology only after it has become widely established in industrialized countries. Costa Rica assembles baseballs for the major leagues; Du Pont turns out aerospace composites. Other economists also point to the billions of dollars wasted on aid that recipients have been unable to integrate into their economies, from tractors without gas stations to steel mills without transport for ore or finished products.

Many preach an enlightened cousin of the trickle-down approach known as appropriate technology. And the information advocates argue that high-tech electronic tools are appropriate—$1,000 microcomputers can do the work that required $1-million mainframes in the industrialized world only a decade ago, and cellular telephones can fend off the lengthy process of laying land lines.

Telecommunications linkages in Ghana, for example, might tie cocoa producers more closely to customers in world markets, allowing them to wheel and deal more rapidly and to increase profits that could eventually flow back to farmers. Service jobs that require manual processing of data are being moved from back offices in high-priced countries to provide jobs for city dwellers in the developing world.

> *"Thanks to informatics, technological shortcuts to development exist today and are within the reach of everyone."*

Moreover, networks could keep academics in developing nations in touch with colleagues in the North. The National Science Foundation is contemplating the funding of U.S. university programs to set up data banks that would let both U.S. and Latin American engineers and scientists share information over the Internet data network.

Dogging all these New Age ideas is the fact that the green revolution and other technological salvations were not exactly unmitigated successes. Agricultural aid definitely improved crop yields for larger farmers with the know-how and capital to buy and use hybrid seed, fertilizer and the other accoutrements of a midwestern U.S. farm. But in India and Pakistan, the poorest subsistence farmers and unlanded laborers were still left on the margins.

Similarly, the benefits of PCs and pocket phones have been confined to a narrow segment of society in developing countries. In India, a more than $200-million software industry has emerged, centered around Bombay and the southern city of Bangalore. Software houses there have crafted, among other applications, programs that let U.S. consumers buy gasoline automatically with a credit card, according to Pyramid Research, a consulting firm in Cambridge, Mass. Yet less than than 1 percent of Indians have phones, compared with 96 percent of Americans. Most branches of the country's largest bank, the State Bank of India, still record transactions by hand.

"In these places, what you may see are enclaves of high-expertise engineers and high-skilled technicians with telecommunications links to wherever they're getting their orders—and a barbed-wire fence between them and the rest of the economy," says W. Brian Arthur, a professor of economics at Stanford University. "The proceeds of what they do may go to some Swiss bank, so they may as well be on the moon."

Africa also has measured some gains from information technology and telecommunications, but a huge gap remains in the ability to meet most countries' needs. The cocoa board in Ghana uses Morse code to communicate with offices all over the country.

The government in Ghana now allows private companies to compete with its own Posts and Telecommunications Corporation, which provides 48,000 telephone lines, according to an article in a journal called *African Technology Forum*. Ghana could accommodate at least 200,000 telephone lines, observers estimate. But at $1,000 to install a handset—more than the nation's average per capita income—demand for privately supplied telephones will remain limited.

Instead, where information technology has become a boon is in the newly industrialized nations, those with the organizational and planning skills to make full use of the technology. Singapore saw a dramatic increase in revenues from software and services during the 1980s. The small nation's advances resulted from an aggressive government policy of building telecommunications networks and promoting a domestic software industry. Other nations—Venezuela and Mexico are two—have witnessed a rise in supply of telecommunications services after privatization of government telephone monopolies.

Indeed, some economists worry that information technology may actually draw jobs away from the developing world, further exacerbating the gap between rich and poor. Just-in-time inventory control methods and computerized manufacturing have become a form of intellectual and organizational capital that outstrips the importance of raw labor costs. In short, Ashoka Mody, a World Bank economist, warns, "the low wages in the developing countries can get lost in the noise of cost accounting."

—Gary Stix and Paul Wallich

Networks

As the diversity of computer applications increases, the burgeoning flow of megabit traffic between machines will be accommodated by wider and smoother highways

by Vinton G. Cerf

A web of glass spans the globe. Through it, brief sparks of light incessantly fly, linking machines chip to chip and people face to face. Ours is the age of information, in which machines have joined humans in the exchange and creation of knowledge.

As the diversity and sophistication of computational machinery have increased, so have the demands on the networks transporting the information and knowledge generated. Pathways must carry scientific and technological data and also provide links for everything from entertainment to complex computational modeling. Systems must be able to sustain communication rates that range from a few characters per second to billions of characters per second, a span that encompasses keyboard character strokes, high-resolution X-ray diffraction imagery and full-motion supercomputer weather simulations. Moreover, multifunction programs may need to talk to other multifunction programs; for example, a user running a word-processing program, a spreadsheet and a desktop graphics program may want to set up a collaborative link with a colleague running similar programs at a distant site.

What is needed to support the burgeoning flow of megabit traffic? What structures are already in place, and what must be built? How should the existing infrastructure—the set of products, services and functions underlying all aspects of networking—be modified? How can users make appropriate choices about the quality of service (speed, accuracy and security, for example)? And how will the system execute these decisions? Such questions fuel current research and development in computer networking. They are made even more challenging by the communications support that multiple applications may need: one can imagine that an unknown arbitrary number of programs running in an arbitrary number of computers may need to communicate with one another at arbitrary times. To meet these requirements, the networking technology must find a way to facilitate the exchange of information among many different computers concurrently. Indeed, the shift from the need to support simple, remote interactive access to a computer to the more difficult task of supporting machine-to-machine interaction has profoundly influenced the development of computer networking technology.

One prerequisite to any successful form of communication is the choice of

NETWORK CONTROL CENTER in the Philadelphia area from which AT&T monitors the flow of voice, data and entertainment traffic typifies the kind of communications system that will require significant enhancement to support growing computerization. Advances in the use of computers in business and professions, expanding consumer services and new entertainment media create the need for changes. The system must be able to handle many classes of traffic, which range in frequency from hundreds of bits per second to billions per second and which bear varying requirements for accuracy and confidentiality.

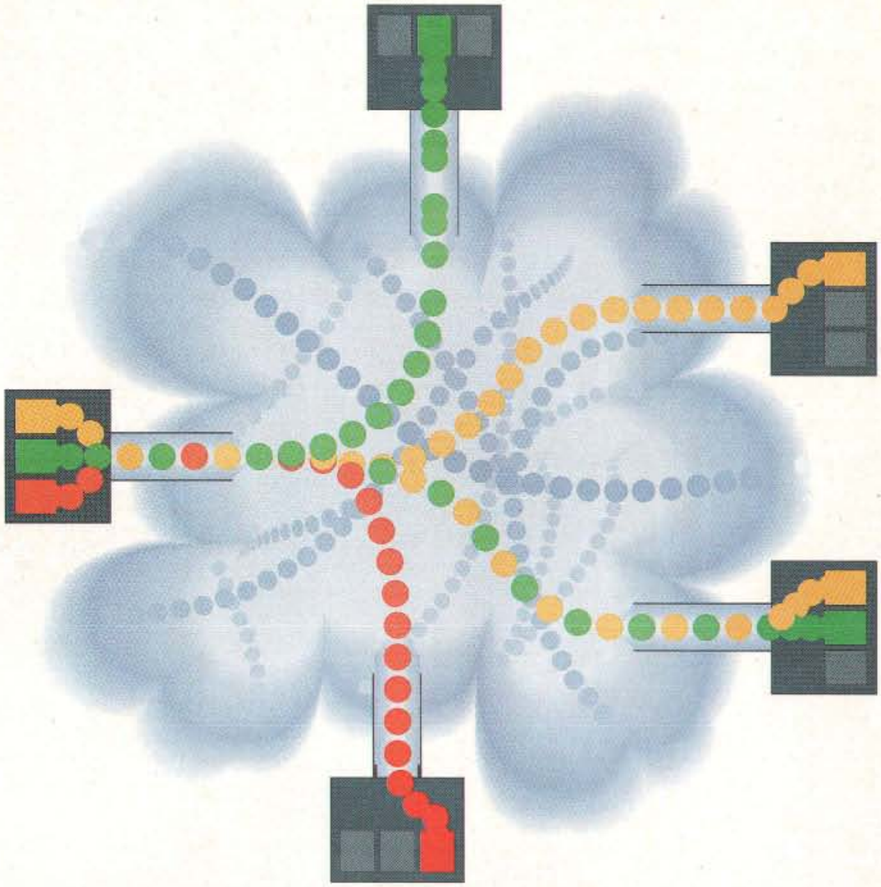
VINTON G. CERF is vice president of the Corporation for National Research Initiatives (CNRI), where he is responsible for the Digital Library System, Electronic Mail Interconnection and Internet research programs. He received his B.S. in mathematics from Stanford University and his Ph.D. in computer science from the University of California, Los Angeles. He taught at Stanford and worked for the Defense Advanced Research Projects Agency and MCI before taking up his present position in 1986.

a common language. In computer networking, it is essential that the communicating programs share conventions for representing the information in digital form and procedures for coordinating communication paths. Like their human counterparts, communicating computers must agree on ground rules for interaction. In personal terms, we might agree to use the postal service to exchange postcards and to use English as our common language.

A computer communications protocol consists of a set of conventions that determine how digital information will be exchanged between programs. The conventions can become quite complex, and so they are often organized in a hierarchy with the most basic agreements at the lowest levels and the more sophisticated and special-purpose agreements toward the top.

Once a language and a set of conventions have been chosen, the next question is how to minimize the number of interfaces ("spigots") required to link a computer to the network. The most effective way is multiplexing: each computer feeds information into a single high-capacity link to the network, to which the other computers are in turn connected [*see illustration on next page*]. Information from each computer is labeled so the network can route information to its proper destination. Ideally, a network can deliver information to all (broadcast), to a subset (multicast) or to only one recipient.

Such manifold computer interactions are often referred to as distributed computing. The term covers a multitude of applications and protocols. A



WIDE-AREA NETWORK moves data among computers through high-speed links. (The network is represented as a cloud, which is the term for it used among computer scientists.) Information packets from several programs running on the same computer may be multiplexed together. In addition, the network can deliver the same message simultaneously to several different addresses, a process known as multicasting.

popular distributed computing concept is called the client-server model. In this model, one computer provides a service, which another computer accesses as a client. For example, one or more computers might be dedicated to storing files of information for all the other computers. Various machines in the network can ask the "file servers" to deliver copies of files on demand. Another set of computers might be dedicated to providing laser-printing services. More generally, a number of computers might be dedicated to providing access to vast quantities of catalogued information in on-line data bases.

Autonomous, interprogram communication is being explored today in an even more general form under the rubric of knowledge robots (Knowbot is a registered trademark of the Corporation for National Research Initiatives), programs that move from machine to machine, possibly cloning themselves. Knowbots support parallel computations at different sites. They communicate with one another, with various servers in the network and with users. In the future, much computer communication could consist of the interactions of Knowbots dispatched to do our bidding in a global landscape of networked computing and information resources.

In any kind of multiplex network, there are two means of relaying information: circuit switching and packet switching. Circuit switching, the familiar technique that links telephones to one another, can make sense whenever two computers need to be connected for a long time to transfer large amounts of information. Its speeds typically range from a few hundred bits per second to a few million. Nonswitched dedicated data circuits can operate at speeds of nearly 45 million bits per second, but the expense of such a system must be justified by a nearly continuous need to transmit data between computers.

Circuit-switched systems suffer from a severe drawback: they often do not prove satisfactory for applications requiring that multiple programs running concurrently in one machine communicate with programs running in different machines. In such cases, it would be necessary to reconfigure the switching system from one computer to the others in turn, a process that takes an inordinate amount of time. Compared with the few microseconds (millionths of a second) or less it takes a computer to finish a computation, setting up and tearing down circuits can take hundreds of milliseconds (thousandths of a second) or more—enough time to complete thousands of computer transactions.

Packet switching avoids this drawback: it is expressly designed to accommodate the bursty, multiprocess communication commonly found in distributed computing environments. Packets, or chunks of data, produced at the originating computer are prefixed with headers containing routing information that identifies the source and destination computers. Small computers, called packet switches, are linked to form a network. Each packet switch examines each header and decides where to send the packet to move it closer to its final destination. When circuit-switched systems cannot accept new traffic, they refuse to set up new circuits. Packet-switched systems, however, exchange this behavior for variations in delay resulting from storing and forwarding packets. Thus, traffic is not refused, only momentarily delayed.

Packet-switched systems offer another bonus. In circuit-switched systems, the sending and receiving computers must usually be able to exchange data at the same rate, such as 9,600 bits per second or 64,000 bits per second. In packet-switched systems, by contrast, the sender might be able to transmit at very high speeds, such as 10 million bits per second, while the receiver might be able to receive only at low speeds, such as 1,200 bits per second. The automatic speed-matching feature of packet-switched systems allows different types of computer systems, such as supercomputers and personal computers, to communicate. Of course, when two very mismatched computers communicate, the faster one must adapt its average rate of transmission to accommodate the slower.

In addition, because the packet switches are programmable computers, they can detect when the trunk lines linking them have failed and select alternate routes for passing packet traffic without disturbing the source and destination computers. A circuit switch can detect similar failures, but the entire circuit has to be reconstructed in order to recover.

Although packet-switched systems can support communication among processes running in many different

Layers of Communication

APPLICATION	Detailed information about data being exchanged
PRESENTATION	Conventions for representing data
SESSION	Management of connections between programs
TRANSPORT	Delivery (reliable or otherwise) of sequences of packets
NETWORK	Format of individual data packets
LINK	Access to and control of transmission medium
PHYSICAL	Medium of transmission (electronic, optical or other)

computers concurrently, they are not without their problems. In some systems, packets may be routed over multiple paths and arrive in a different order from the one in which they were sent. The receiving computer has to make sure packets are put back in the original sequence (in some systems this service is provided by the receiving packet switch). A disorderly arrival can also occur if a packet is retransmitted because of a detected error.

Another problem is congestion. The packet flow may overtax the storage and forwarding capacity of the switches. If that happens, the switches may be forced to discard packets that cannot be stored or delivered. Detection and prevention of congestion is an important area of packet-switching research.

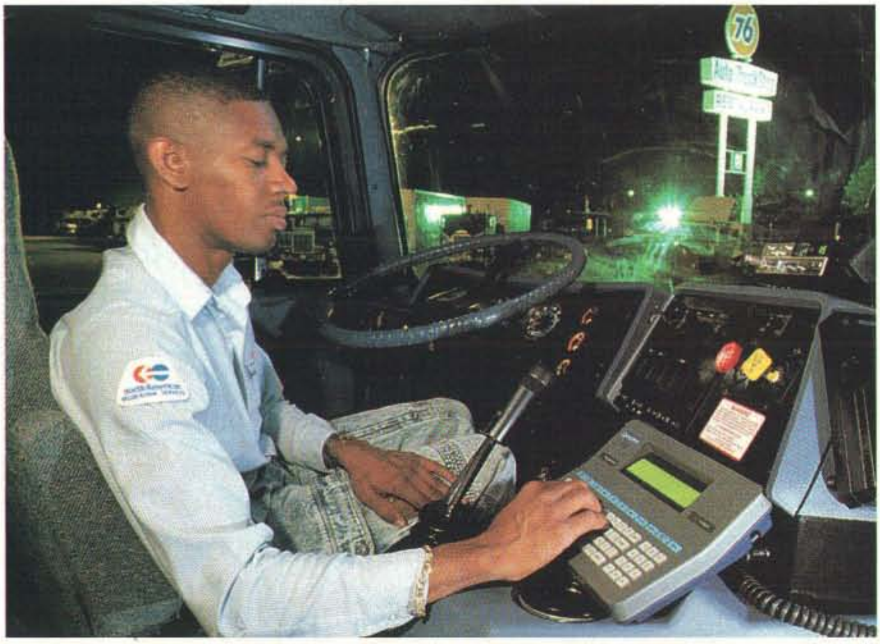
To avoid disorderly arrivals, some packet-switched systems offer a virtual circuit service. In this system the network appears to provide a dedicated circuit to deliver information from one computer to another. In reality, the circuit is virtual, delivering packets in order but with variable delay. Usually the traffic routes are fixed and change only if a failure takes place. If the traffic on each virtual circuit is variable—if the packets emerge from the source in a bursty fashion—then route selection to minimize congestion can still present a challenge. An international standard for virtual circuit packet switching has been codified by the Consultative Committee for International Telephony and Telegraphy (CCITT), an arm of the International Telecommunications Union (ITU). The ITU, which is affiliated with the United Nations, is an international treaty organization.

One of the first packet-switched systems was developed in 1969 under contract to the U.S. Defense Advanced Research Projects Agency (DARPA). Called the ARPANET, the system used minicomputers as packet switches and dedicated 50-kilobit-per-second telephone lines to connect them. Similar projects were also started in other countries, notably the U.K. (the National Physical Laboratory) and France (Institut National Recherche d'Informatique et d'Automatique). Since then, a great many private and public packet-switched networks have been deployed. They typically operate using trunk lines at 56,000, 64,000 or 1.5 to 2 million bits per second.

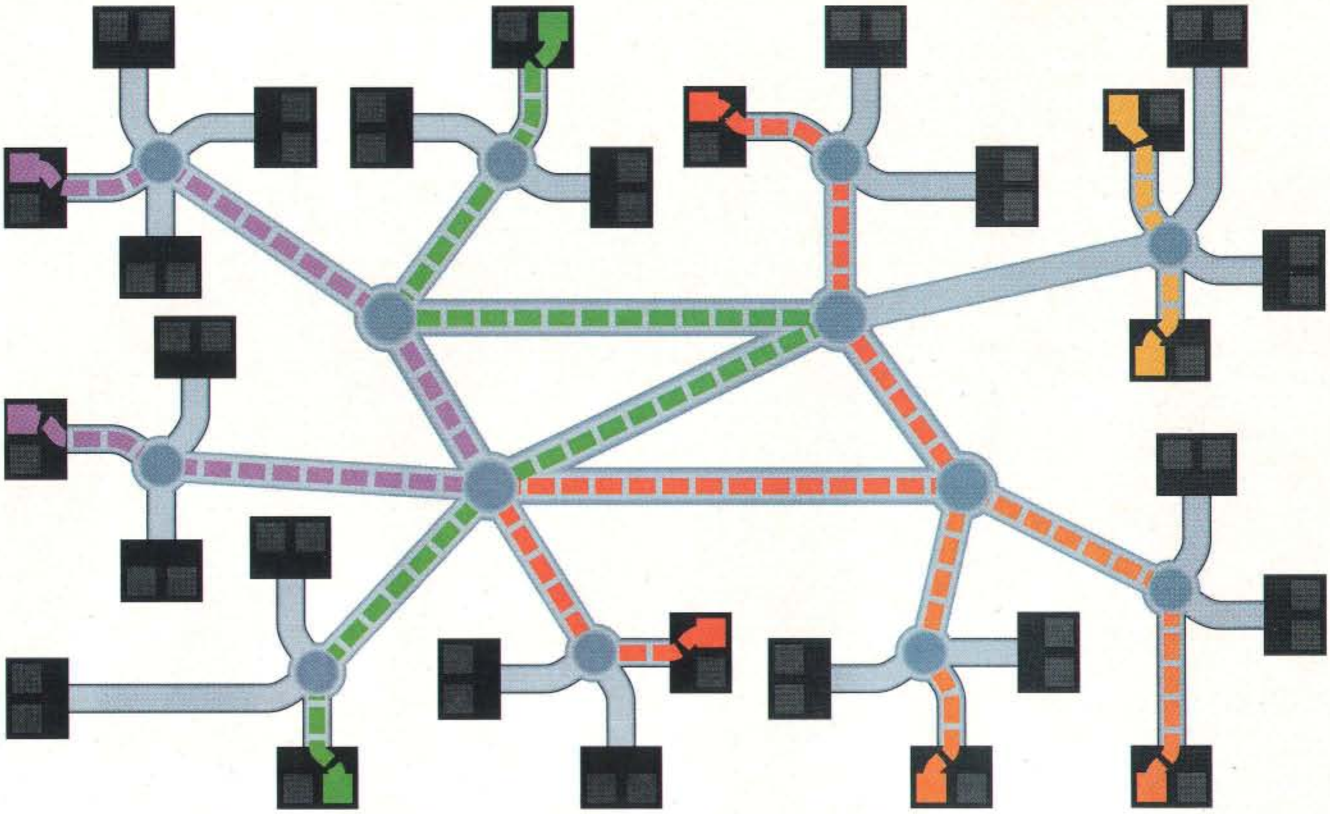
Concurrent with its research on the ARPANET, DARPA also explored packet-switching methods in mobile radio systems and synchronous satellite systems. The ARPANET and its satellite-based counterpart were early examples of so-called wide-area networks (WANs).

Packet-switched systems are used on a variety of transmission media, including local-area networks, metropolitan-area networks, integrated-services digital networks (ISDNs) and networks that operate at gigabit (one billion bits) speeds. One of the first local-area networks to use packet switching was developed in the early 1970s by the Xerox Palo Alto Research Center [see "Networks for Advanced Computing," by Robert E. Kahn; SCIENTIFIC AMERICAN, October 1987]. The system, called Ethernet, is still used today. An Ethernet network transmits signals, which are heard by all receivers (in a broadcast), by coaxial cable over distances of a kilometer or two. Initially operating at three million bits per second, modern Ethernets now work at 10 million bits per second, taking 100 nanoseconds (billionths of a second) to send each bit. Electric signals propagate on the coaxial cable at roughly half the speed of light, or 150,000 kilometers per second. Consequently, one bit can propagate through a 1.5-kilometer network in 10 microseconds.

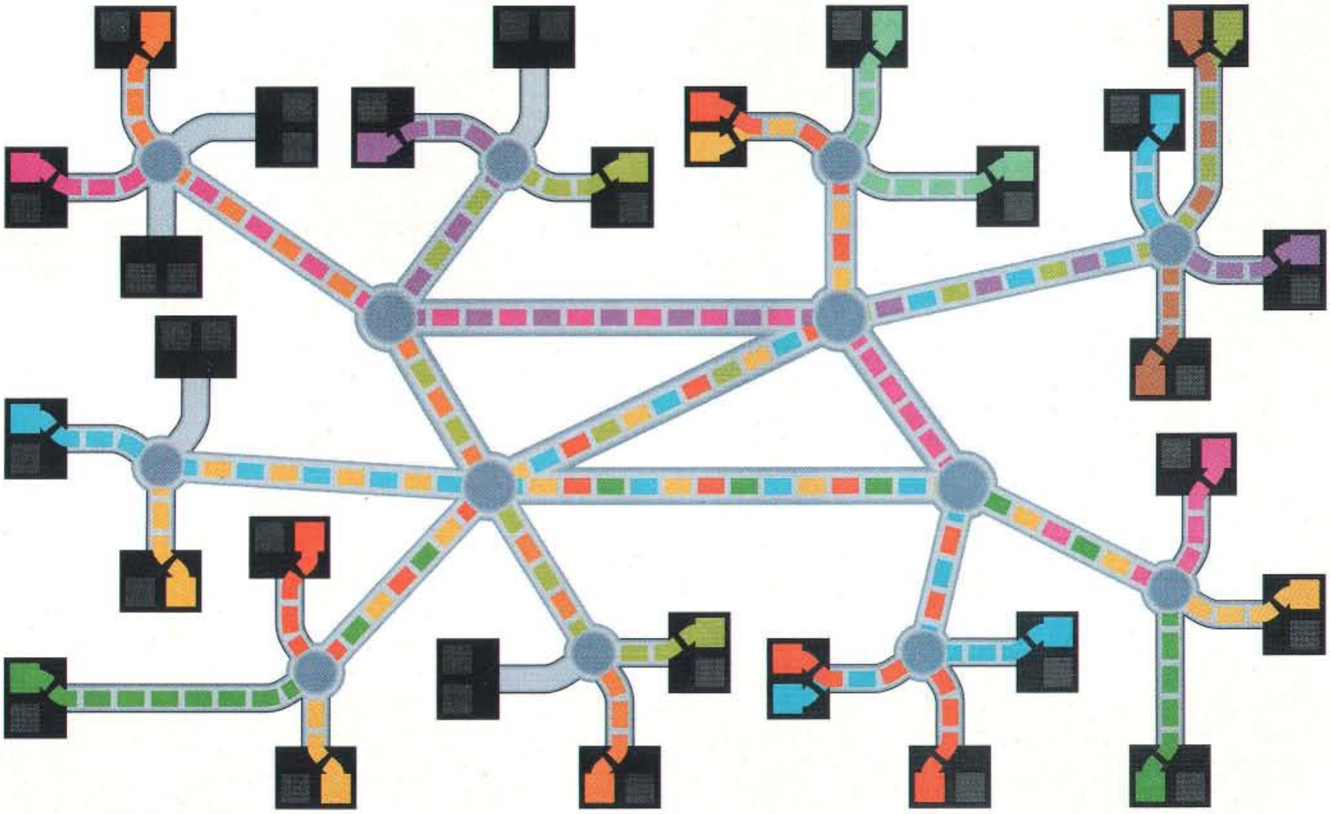
Each transmitter on the Ethernet listens before it transmits. If another transmission is detected, it ceases and waits for a random interval before retrying. This method of channel sharing is called carrier-sense multiple access (CSMA). The Ethernet design cleverly in-



WIRELESS NETWORKS will permit mobile workstations to communicate with the global network. The truck driver in the photograph is using a satellite-based system developed by Qualcomm, Inc., in San Diego, Calif. The system provides instant two-way communication between the trucker and the dispatch center, even on a cross-country run, allowing for unscheduled pickups that reduce "deadhead" miles.



CIRCUIT SWITCHING creates an end-to-end path through the network for data flowing between two computers. This technique, based on methods developed for connecting telephone calls, simplifies data transmission. Yet it is not efficient, because setting up the connection typically takes far more time than transmitting the desired message.



PACKET SWITCHING reduces data transmission overhead by attaching codes (*shown as different colors*) to each item of data sent through the network. The codes identify the source and the destination of each item, so that no end-to-end connection needs to be established. Thus, packets from many different computers can travel easily over the same network links.

cluded a test for competing transmissions *during* transmission so that a collision at the start of sending would cause all colliding senders to stop and wait for a random length of time. The deleterious effect of simultaneous transmissions was thus minimized. If the maximum span of an Ethernet is kept under 1.5 kilometers, collisions, or conflicting transmissions, will be quickly detected and will not consume much capacity on the network. A larger network, however, would suffer a much higher incidence of collision.

A development made in parallel with Ethernet was based on the idea of a ring of computers in which a token, or a short series of bits, is passed from computer to computer. A computer receiving the token is free to transmit one or more packets; other computers must wait until they receive the token to transmit. Token-based systems, which can also broadcast and multicast messages, generally operate at four million to 16 million bits per second.

More recently, packet technologies based on fiber-optic transmission have been developed. They operate at higher speeds and are better suited to larger, metropolitan-area networks. One example is the Fiber Distributed Data Interface (FDDI), which operates at 100 million bits per second and uses a token-based approach to sharing the fiber capacity. The system is organized in a dual ring, so that if a section fails, full operation can be rapidly recovered [*see illustration on this page*]. Token systems can span larger distances than broadcast Ethernets, but the price paid is a longer delay to access the ring. Care must be taken not to introduce more than one token on the ring at a time, although some variations permit multiple tokens to circulate.

A recent development is a fiber-based network technology called Distributed Queue Dual Bus (DQDB) networking. In this design, which works over tens of kilometers or more, nodes are connected to two different fibers, one for each direction. At the originating ends of each fiber, a special node sends out empty packets. The first node that has a packet of data to send fills the empty packet and sends it on its way. When a node has data to send to another node, it sends a request indicator on packets flowing on both fibers. In this way, upstream nodes learn that a downstream node has a packet to send. As empty packets appear, each upstream node allows them to propagate downstream to the node that needs to send data. Each node thus keeps track of the queue of packets awaiting transmission downstream.

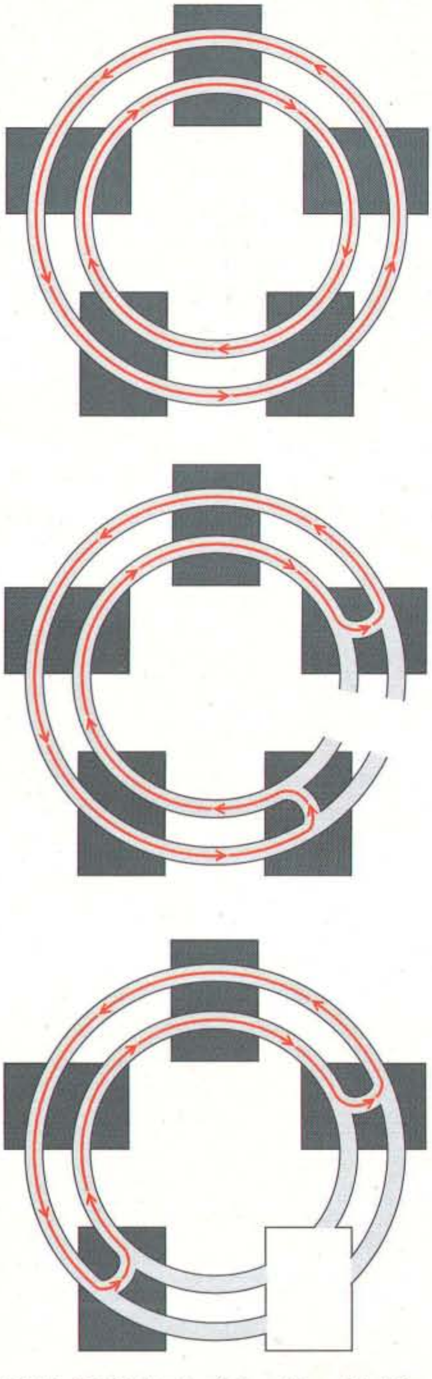
The DQDB technology, which was developed by the University of Western Australia in the mid- to late 1980s, is undergoing trials by local exchange carriers in the U.S. and by telephone administrators in other countries. Because the service is expected to operate at speeds up to 600 million bits per second, it may figure prominently in metropolitan-area networks.

Another recent networking technology is Frame Relay. It is similar in nature to the virtual circuit system, except that the virtual circuits are determined at the time subscribers are connected to the system. No error checking is done as the packets move from switch to switch; there is only end-to-end error checking and retransmission, thereby reducing the delay through the network. Frame Relay, expected to operate in the range of 64,000 to 45 million bits per second, can be used in local-, metropolitan- or wide-area networks.

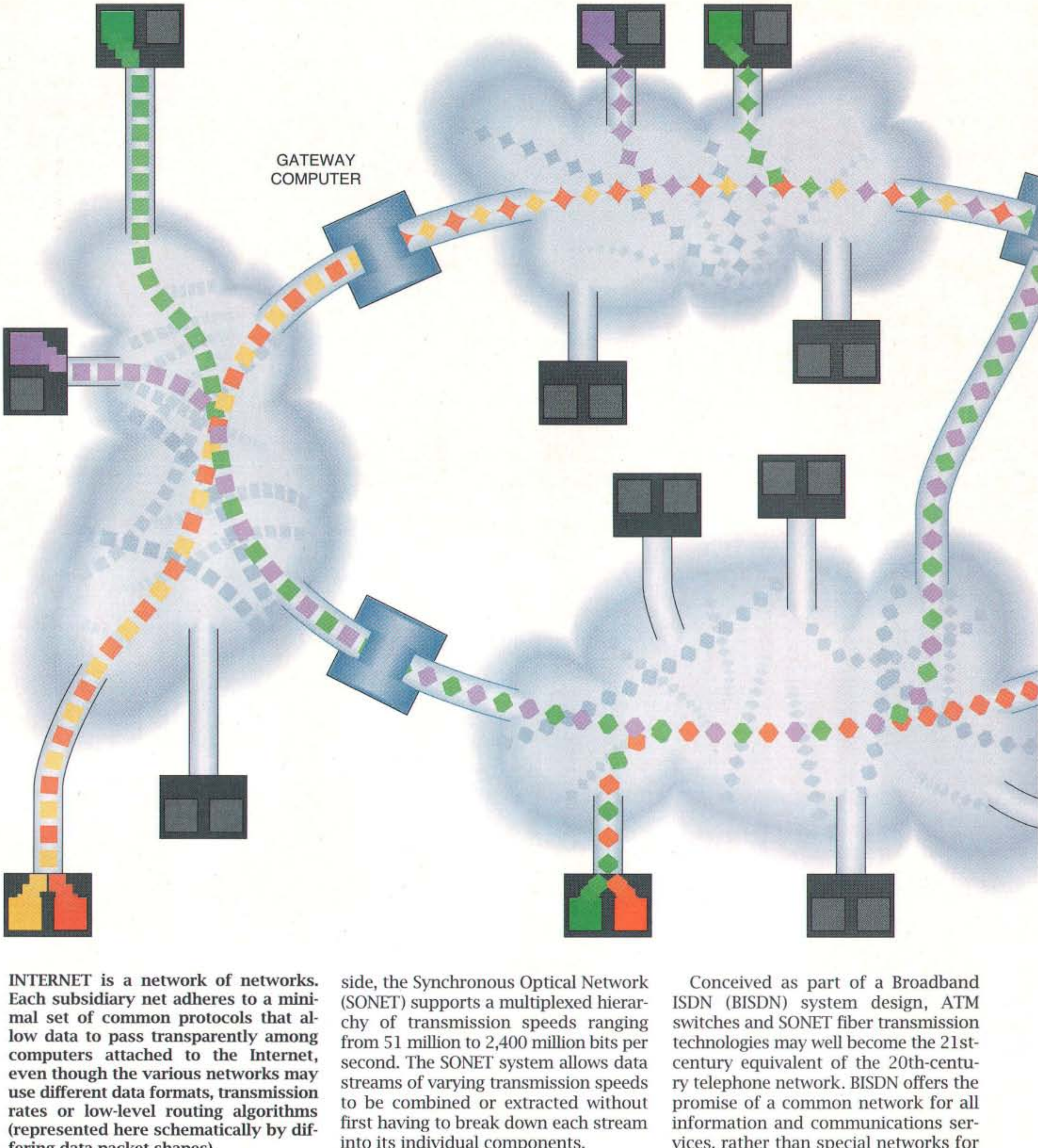
Over the past 15 years, the telephone carriers have been developing a wide-area digital communications technology, the ISDN, that would permit both voice and data to be carried in digital form across an international, switched digital backbone. ISDN offers two kinds of access to the digital transmission medium. The first, called Basic Rate Interface (BRI), gives a subscriber two bearer (B) channels that operate at 64,000 bits per second each and one data (D) and signaling channel that operates at 16,000 bits per second. The D channel is used to signal the network where to connect each of the B channels. The second is called Primary Rate Interface and operates at 1.5 to 2 million bits per second, providing 23 to 30 B channels.

Unfortunately, the service has not been widely deployed, nor has there been a great demand for it. One reason for the cool reception may be that it has been possible to use the voice network to support data communications at speeds up to 19,200 bits per second by employing a device called a modem (modulator-demodulator). The modem turns digital binary signals into modulated sound signals that can be propagated anywhere in the voice network. The difference between 19,200 and 64,000 bits per second is a factor of three, but apparently this difference has not been attractive enough to overcome the cost barrier of having to purchase special equipment to interface with the ISDN network.

An exciting recent development is the emergence of the gigabit-speed network. Local networks operating at speeds in the range of a billion bits per second are now being designed by running parallel connections between computers. For example, a ribbon cable containing 64 conductors could sustain an effective data rate of a billion bits per second if each cable carried 16 million bits per second. (Cables carrying 10 to 20 million bits per second are commonplace.) One such system is marketed by Ultra Network Technologies for



DUAL RING (*top*) of the Fiber Distributed Data Interface network helps to safeguard connections against failure. If one of the FDDI optical fibers fails (*middle*) or if a network node ceases operation (*bottom*), the adjacent nodes can readily reconfigure the system to restore full operation.



INTERNET is a network of networks. Each subsidiary net adheres to a minimal set of common protocols that allow data to pass transparently among computers attached to the Internet, even though the various networks may use different data formats, transmission rates or low-level routing algorithms (represented here schematically by differing data packet shapes).

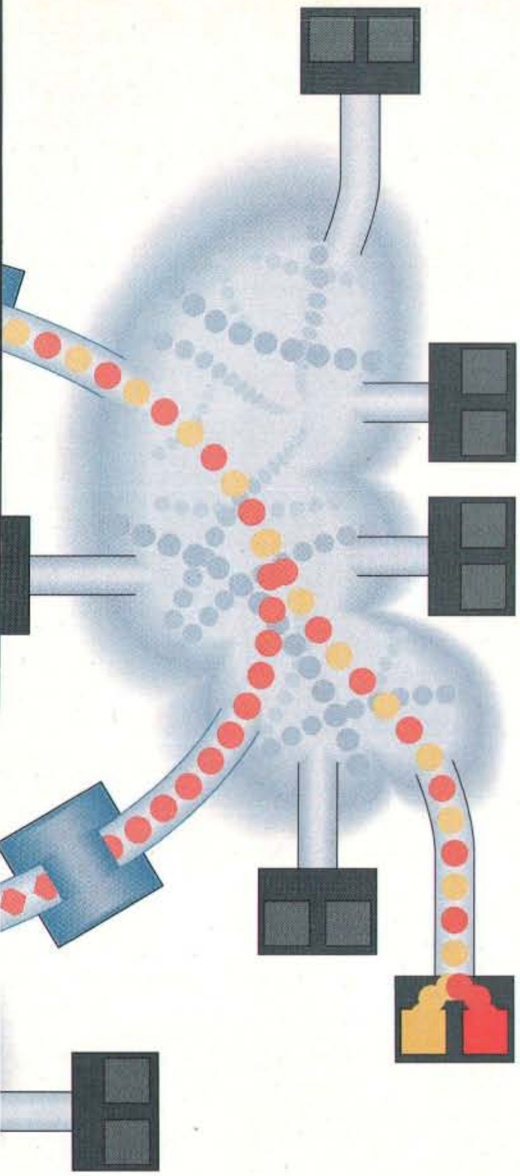
the purpose of linking supercomputers.

For the past few years, fiber-based technologies have been emerging from the laboratory and making their appearance in experimental wide-area settings. These transmission and switching technologies are designed to operate at speeds in excess of a billion bits per second. On the transmission side, the Synchronous Optical Network (SONET) supports a multiplexed hierarchy of transmission speeds ranging from 51 million to 2,400 million bits per second. The SONET system allows data streams of varying transmission speeds to be combined or extracted without first having to break down each stream into its individual components.

Complementing this new transmission technology is a fast packet technique, Asynchronous Transfer Mode (ATM) switching, which can switch short packets called cells at extremely high rates. Cells, which contain up to 48 octets (an octet is eight bits) of data and five octets of addressing and control information, can carry digitized voice, arbitrary data and even digitized video streams.

Conceived as part of a Broadband ISDN (BISDN) system design, ATM switches and SONET fiber transmission technologies may well become the 21st-century equivalent of the 20th-century telephone network. BISDN offers the promise of a common network for all information and communications services, rather than special networks for different services, such as voice, data and video. Already several experimental test-beds, sponsored by DARPA and the National Science Foundation, are exploring applications and architectures for building wide-area gigabit networks.

In addition to the high-speed network developments under way, researchers are working on wireless digital networks that will permit mobile workstations to be part of the global



network. Wireless local-area networks capable of transmitting 10 megabits per second within a room or building are already available, and experimental systems offer far greater bandwidth.

The wireless component of the global net will bear some resemblance to a cellular telephone network—computers carried in trucks, ships or briefcases will be able to maintain connections and send and receive data wherever they go—but current cellular technology is not appropriate for wireless data transmission. First, cellular telephones rely on analog broadcast techniques, so that using them to move digital data would be inherently inefficient. Digital information would have to be converted to analog form for wireless transmission, much as computers now employ modems to send data over ordinary telephone lines.

Second, existing cellular networks are already unable to process the full load of calls that their users desire, much less additional data traffic. A single telephone "cell," typically several miles across, can handle only 59 connections at a time. A few dozen cellular modems could easily block out almost all other callers.

Third, even under the best of circumstances, the frequencies assigned to the cellular telephone network can provide data transmission rates of no more than 100,000 bits per second—insufficient for many potential mobile applications. A wireless complement to existing and future global networks will require new technology, massive capital investment and regulatory changes to allocate sufficient space for data transmission in the electromagnetic spectrum.

Although it may not be apparent, the various technologies I have described can be organized into an architectural hierarchy. Such a conceptual base helps in designing new computer communications technology. Since any computer communications system is based on its protocols, it is not surprising that the conceptual structure is a protocol hierarchy. From the bottom up, the layers can be labeled physical, link, network, transport, session, presentation and application [*see top illustration on page 47*].

The physical layer has to do with the actual medium of electronic, radio frequency or optical transmission and the way in which bits are signaled on the medium. The link layer determines how sequences of bits are framed into chunks. The network layer deals with packet communication and is typically the lowest level at which computer programs can communicate.

The transport layer is the first one at which end-to-end flow and congestion control between communicating programs are regulated. Some applications require that data be delivered in sequence, with high reliability. Others require only that data be delivered expeditiously—some information can be lost. For example, voice or video packets can be made to work if the delays are minimal and the arrival time between packets is short. A lost speech or video packet might introduce a brief gap, but if gaps are infrequent, listeners and viewers can ignore the problem. On the other hand, a file containing a computer program must arrive intact, and an accurate and sequenced delivery is essential.

The layers above transport are closer to the applications and often reflect their needs. Associations between communicating programs are established at the session layer. Conventions for representing information to be exchanged are determined at the presentation layer. Spanning all the layers is the management layer, since management pervades all aspects of networking, from the lowest level to the highest.

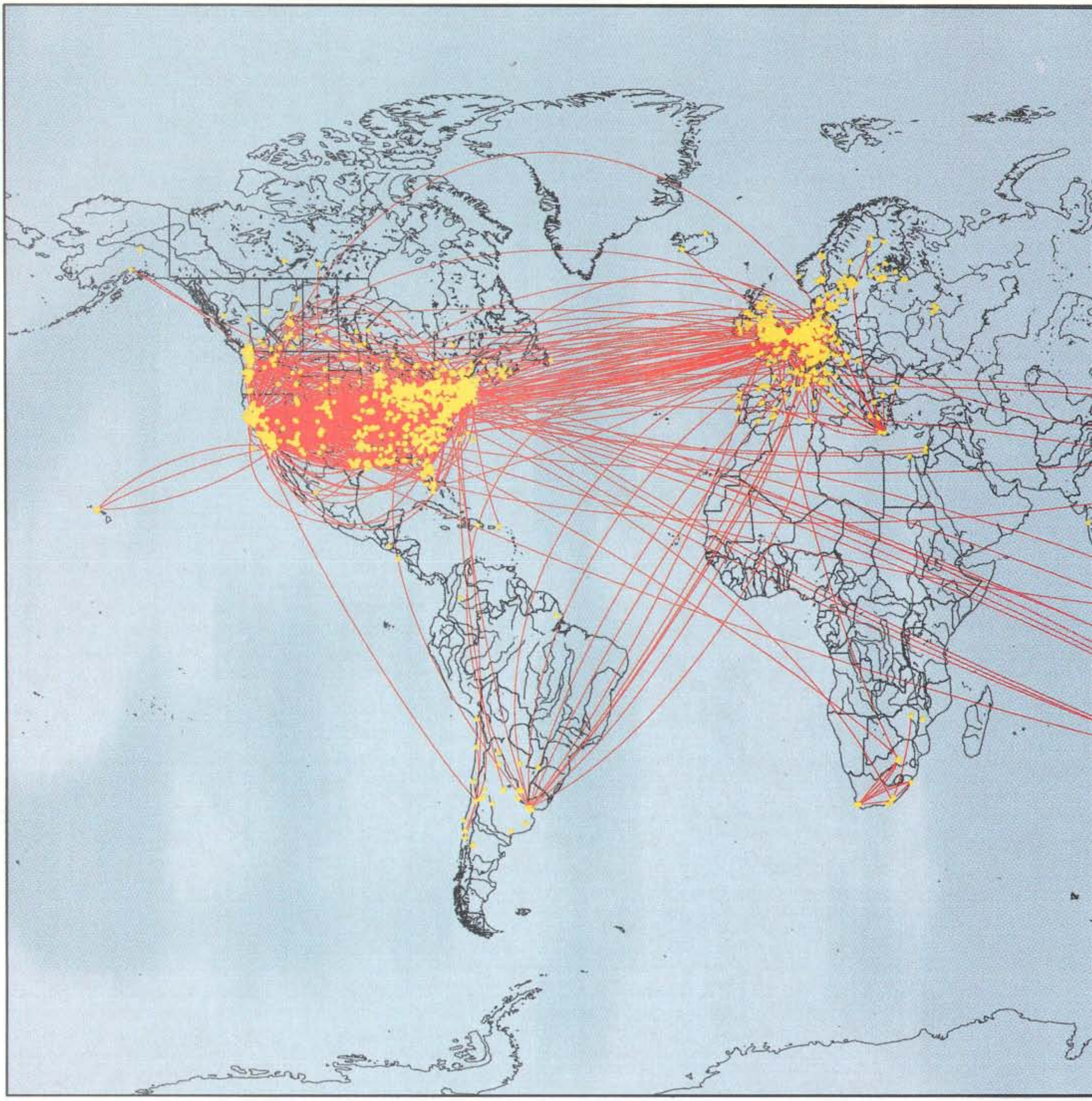
In nearly all the discussion so far, I have focused on the first three protocol layers—physical, link and network. I now move to the higher layers—and straight into the important concept of networking the networks, a procedure called network interconnection, or internetting.

Internetting was first explored in the early 1970s by DARPA. The agency sought ways to interconnect different kinds of packet networks so that the computers using them could communicate without concern for which kind of and how many networks made the interconnections. Special processors called gateways were developed to connect two or more networks and to pass packet traffic from one network to the next. Gateways were also responsible for dealing with network differences, such as speed, maximum packet length and error rates [*see illustration on these two pages*].

An internet gives each of its users an address and defines a standard packet format. To send information, a computer creates packets including the source and destination addresses and encapsulates the packets in the format required by the underlying network. The computer then routes the packets to the appropriate gateway or host for further processing, and this series of operations is repeated until the packets reach their final destination.

Many of the same problems that arise in ordinary networks arise in internets. Gateways need routing algorithms so that they can determine the topology of the relevant parts of the internet and decide where internet packets should be sent. Changes in topology caused by network failures and gateway crashes must be accommodated. Control of flow and congestion in an internet pose as much of a challenge as they do in lower-level networks.

In the boiling ferment of modern telecommunications technology, a critical challenge is to determine how the internetting architecture developed over the past 15 years will have to change to adapt to the emerging gigabit-speed ATM, BISDN and SONET technologies of the 1990s. DARPA and the National Science Foundation are sponsoring

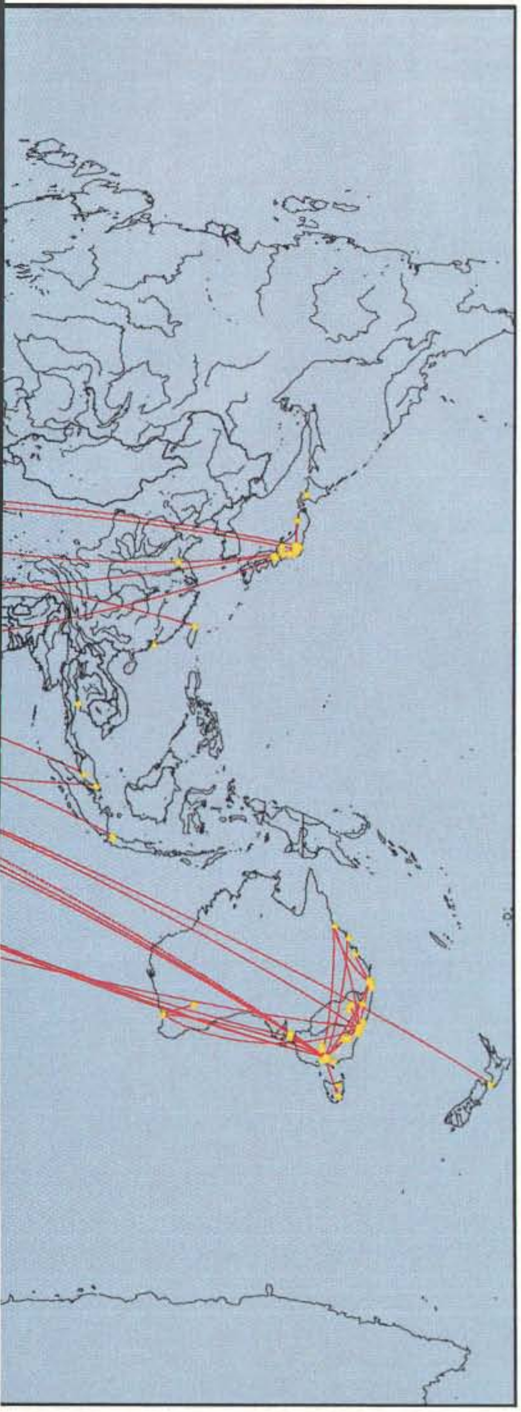


USENET, a worldwide, voluntary member network, has approximately 37,000 nodes (about one tenth the size of the Internet). Accessing electronic mail and other network services requires only finding another Usenet site that is willing to supply a connection.

a major test-bed program to examine this problem through the exploration of experimental applications requiring such superhigh-speed networking. Supercomputer simulation imagery ("visualization"), multiple supercomputer computations and medical and geophysical image generation are all being used to test protocol designs, architectural alternatives and programming environments.

A number of proprietary internetting technologies have been developed by Xerox (the Xerox Network System), Digital Equipment Corporation (DECnet Network Architecture, or DNA) and IBM (Systems Network Architecture and Systems Network Interconnection). DARPA initiated one of the largest open-network systems, the Internet. It operates in 26 countries, comprises more than 5,000 networks and supports several million users on more than 300,000 computers in several thousand organizations. In the U.S., the Internet system has strong support from DARPA, the National Science Foundation, the National Aeronautics and Space Administration and the Department of Energy.

Operating a complex and large-scale computer network or collection of computer networks is a complicated enterprise. As the number of devices involved in the system increases, the system's complexity grows exponentially. Detecting and repairing software, machine and communications



link failures are extremely difficult. As might be expected, network management is a major target for research and development.

A crucial area of network management concerns the security of the system at all levels. Users of remote resources must be satisfactorily identified—typically, by means of a password. Unfortunately, this approach is weak, in part because passwords are not chosen well (they are often last names, names of spouses, license plates and birth dates, for example) and in part because passwords are carried without special protection across the network and thus may be observed by those technically equipped to do so.

The need for security and, in particular, authentication arises at all levels of the protocol hierarchy. At the top level, users may want assurance that electronic mail is coming from the person purported to be the sender; processors may need to know, for accounting purposes, which other systems are consuming resources or, for access-control purposes, which systems are accessing information. In financial transactions, it is critical to assure integrity (that is, to make sure there has been no tampering with a message). For example, one would like to be certain that a deposit is not surreptitiously transferred to another account. In business transactions in which orders are placed, it would be helpful if it were difficult to repudiate confirmed orders.

At lower layers, gateways and routers need to know that control commands are coming from legitimate network management stations. In addition, sometimes information exchanged on the network, such as medical records and electronic mail, needs to be kept confidential. In other cases, it is critical that information such as financial transactions, business orders, network-control information and accounting records not be altered in transit.

Digital cryptography can often satisfy these various requirements. The National Institute of Standards and Technology sponsored the development of a Data Encryption Standard (DES) in the mid-1970s for commercial and government users who did not require military-level cryptography. During the same period, the concept of public key cryptography (PKC) was developed.

In conventional cryptographic systems, a single key encrypts and decrypts messages between parties who wish to keep their communication private and who want assurance that only parties who hold the keys are able to communicate. The DES algorithm is a conventional system: any party holding the key can encrypt or decrypt messages. Such a system is sometimes called a symmetric keying system, since all parties use the same keys, and the same cryptographic algorithm serves for both encryption and decryption.

Public key systems, in contrast, use a pair of keys [see "The Mathematics of Public-Key Cryptography," by Martin E. Hellman; SCIENTIFIC AMERICAN, August 1979]. Messages encrypted in one key can be decrypted only with the other key. It does not matter which key is used to encrypt, only that the other key must be used to decrypt. This form of cryptography is sometimes called an asymmetric keying system.

Typically a user of a public keying system will keep one key private and publicize the other (hence the term "public key system"). To transmit a confidential message to a recipient, the sender encrypts the message in the recipient's public key. Only the recipient can decrypt the message, since only the secret key is useful for that purpose.

An interesting twist to public key cryptography is the digital signature. To "sign" a message, the sender encrypts the message in the secret key. The recipient is given the encrypted message and the advice that it came from the sender. The recipient looks up the sender's public key and uses it to decrypt the message. If the decryption is successful, the recipient knows that the message came from the sender, since only the sender has the secret key that matches the public one. Clearly, digital signatures can be exploited for various transactions (including network management exchanges) in which it is important to verify the source of the message. More important, they allow commerce that would otherwise suffer from serious threats of abuse.

In its book on computer and network security, *Computers at Risk*, the National Research Council opens the first chapter with the alarm, "We are at risk." This is a fair assessment for heavily networked computer systems operating in many countries and with large numbers of programs running concurrently. Technologies for ameliorating that risk are emerging from the research community. If we are at risk, we are also forewarned and, increasingly, forearmed with powerful methods for protecting the system.

FURTHER READING

NEW DIRECTIONS IN CRYPTOGRAPHY. Whitfield Diffie and Martin E. Hellman in *IEEE Transactions on Information Theory,* Vol. IT22, No. 6, pages 644–654; November 1976.

PACKET COMMUNICATION NETWORKS. Special issue of *IEEE Proceedings,* Vol. 66, No. 11; November 1978.

THE MATRIX: COMPUTER NETWORKS AND CONFERENCING SYSTEMS WORLDWIDE. John S. Quarterman. Digital Press, 1989.

COMPUTERS AT RISK: SAFE COMPUTING IN THE INFORMATION AGE. National Research Council System Security Study Committee Staff. National Academy Press, 1990.

COMPUTERS UNDER ATTACK: INTRUDERS, WORMS AND VIRUSES. Edited by Peter J. Denning. Addison-Wesley, 1990.

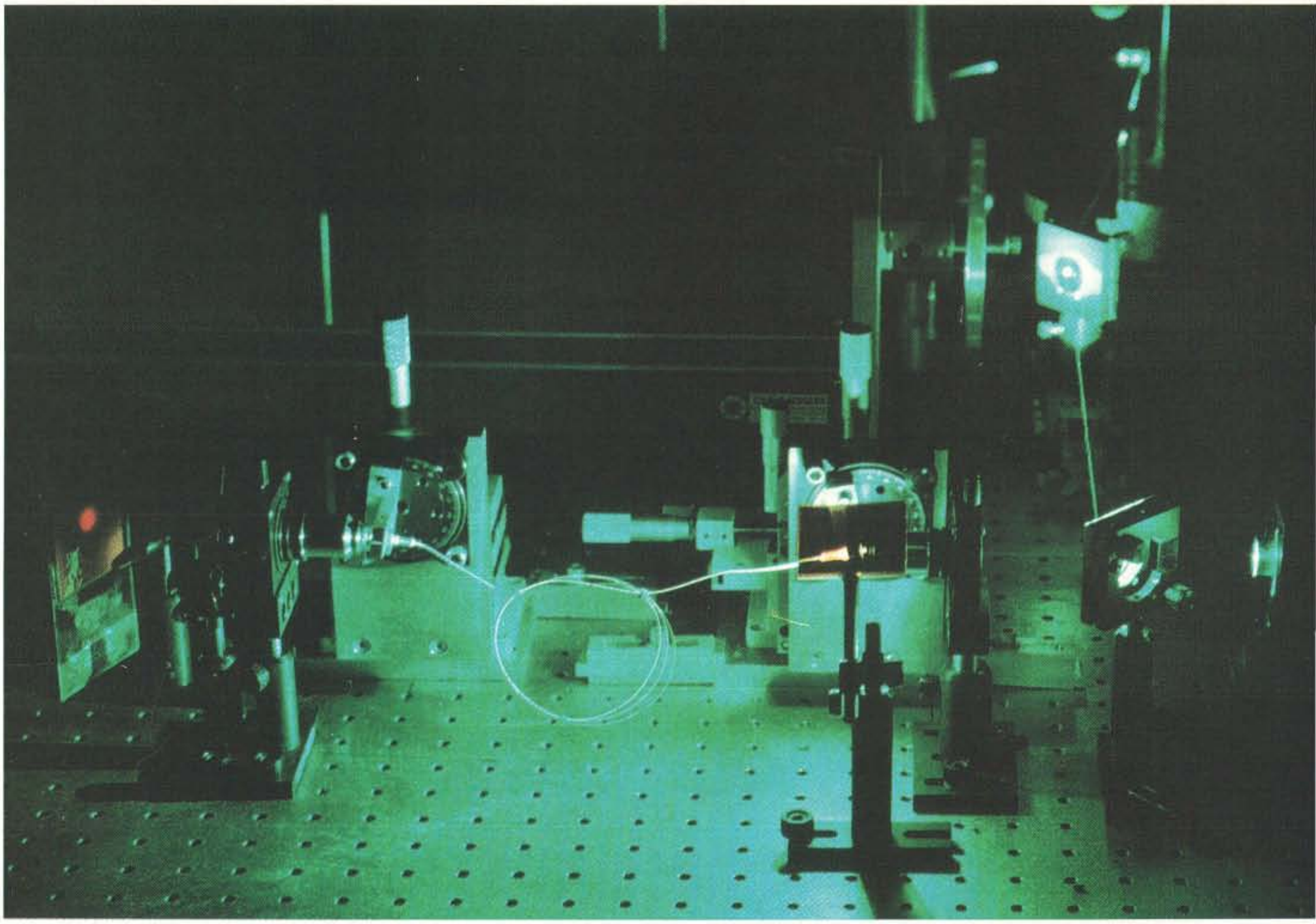
Lightwave Communications: The Fifth Generation

Optical fibers doped with erbium and powered by tiny laser chips are revolutionizing the way signals are regenerated for transcontinental communications and for fast data transmission over fiber-optic networks

by Emmanuel Desurvire

Sometimes innovations emerge when existing technologies or old ideas are applied in unconventional or original ways. Some five years ago many researchers predicted that lightwave communications systems were approaching their peak performance. Ignoring such opinions, a team of researchers reexamined a 20-year-old technology: optical fibers made from silica glass and traces of a rare-earth element. Such fibers can amplify light signals when they are energized by infrared radiation.

Within three years investigators, including myself, had developed a novel type of optical amplifier that is based on a fiber doped with the rare-earth element erbium and is powered by a tiny, efficient radiation source called a laser diode chip. When this fiber amplifier was integrated into communications systems, it increased the transmission capacities of the systems by a factor of 100. In principle, billions of bits of information can be transmitted in a fraction of a second through an uninterrupted cable that stretches a quarter of the way around the planet. In the middle of this decade AT&T Bell



Reprinted from the January 1992 issue

Laboratories and Kokusai Denshin Denwa (KDD) plan to deploy such a cable across the Pacific Ocean, and many other companies are likely to do the same. Each cable will have the capability to carry 500,000 phone calls simultaneously—more than 12 times the capacity of existing transoceanic systems.

Erbium-doped fibers eliminate the need for complicated devices called repeaters, which are used in conventional systems to regenerate weakened signals. Fiber amplifiers can boost signals by greater factors than repeaters, and they can transmit data at higher rates.

The erbium-doped fiber amplifiers will play a major role not only in long-distance communications but also in local-area networks. Fiber amplifiers are ideal for networks that carry vast quantities of information to thousands of users. Such networks could bring high-definition television and video teleconferencing to every home and business.

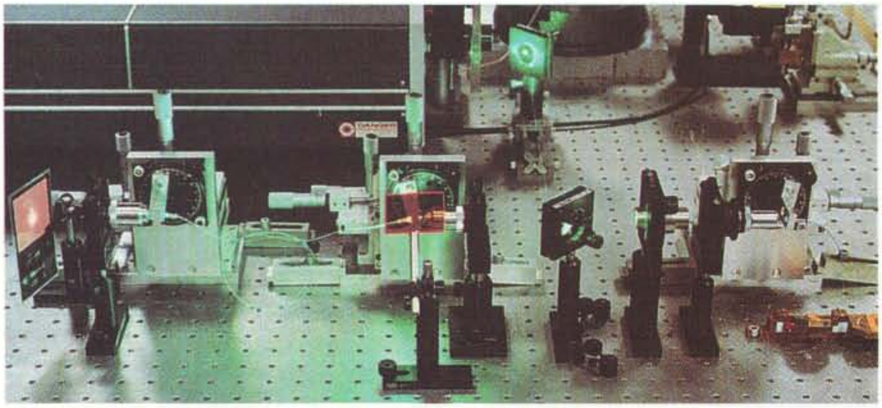
Fiber amplifiers should also be an integral part of future communications systems. They may be incorporated in systems that convey information as a change in the phase and frequency of light rather than as a modulation in intensity as all current systems do. Fiber amplifiers may be important for systems in which information is carried by solitons. These short light pulses can in theory propagate indefinitely in a defect-free optical fiber. If a practical system emerges from any of these concepts, it should greatly exceed the performance of existing systems.

Since 1975 the technology of lightwave communications has progressed, by any standard, at an extremely rapid pace. The technology has evolved as scientists have worked persistently to invent systems that transmit more information at faster rates over longer distances. Innovative systems must also prove practical and cost-effective. During the past 20 years, researchers have launched five generations of lightwave communications systems, each representing a major advance in the technology.

When the first generation of lightwave communications systems was introduced around 1975, it could transmit far more information than could systems carrying electrical signals through copper wires. First-generation systems and those that followed consist of the same basic components: encoding and decoding devices, a transmitter, silica fibers, repeaters and a receiver. The encoding device first converts information into an electrical signal. The transmitter then converts the signal into light. The transmitter, at least in 1975, was a light-emitting diode that produces infrared radiation at a wavelength of 0.87 micron. The diodes respond to changes in the electrical signal by emitting light of various intensities.

In all generations of lightwave systems, light from the transmitter is carried through fibers made of silica glass. A fiber consists of a core, which carries the light, and a cladding, which guides light through the core. As the light signal travels along the fiber, it broadens and weakens. The signal disperses, in part, because radiation of different frequencies propagates at different speeds through the core. The signal attenuates largely because defects or impurities in the glass absorb or scatter light.

A system can compensate for attenuation with repeaters. These devices are placed between lengths of fiber to detect, amplify and reemit the signal.



ERBIUM-DOPED OPTICAL FIBER (the green-glowing coil at the left) will be one of the key components in the next generation of lightwave communications systems. To demonstrate the capabilities of the erbium-doped fiber as an amplifier, the author and his colleagues, then at AT&T Bell Laboratories, built the apparatus shown here in the light (*above*) and dark (*left*). The fiber amplifies infrared light signals when it is energized by green laser light. The erbium atoms in the fiber absorb the green light and jump to a high-energy state. As an infrared light signal passes through the fiber, the erbium atoms transfer their energy to the signal by a process known as stimulated emission. Red light is used to indicate the path of the signal. The signal travels from the right of the apparatus, propagates through the erbium-doped fiber and emerges on the left, where the amplification of the signal is measured.

EMMANUEL DESURVIRE is associate professor of electrical engineering at Columbia University. He works at Columbia's Center for Telecommunications Research. In 1981 he received a diploma of advanced studies (DEA) in theoretical physics from the University of Paris VI. After investigating Raman fiber amplifiers for two years at Thomson-CSF in Orsay, he obtained a doctoral degree from the University of Nice. For the next two years, at Stanford University, he conducted research on active fiber devices. From 1986 to 1990 at AT&T Bell Laboratories, he worked on erbium-doped fiber amplifiers.

Eventually the signal reaches the receiver, which transforms the light back into an electrical signal. Finally, the decoding device translates the electrical signals into a form a telephone or computer can interpret.

The success of each generation of lightwave communications systems ultimately depends on two quantities: how much information can be transmitted in a second through the system and how far signals can travel through a fiber before the signal must be regenerated. The success of the system can therefore be expressed as the transmission capacity, which is defined as the highest bit rate of the system times the maximum transmission length. By 1978, one billion bits (a gigabit) could be transmitted each second through a system 10 kilometers long. The transmission capacity was therefore about 10 gigabit-kilometers per second.

During the next three years, lightwave communications systems reached nearly 100 gigabit-kilometers per second as the second generation emerged. Scientists had improved the technology in two ways. First, by reducing the size of the core of the fiber, they created "single-mode" fibers, which force the light signals to travel at a nearly uniform velocity. This advance greatly reduced dispersion. Second, they developed transmitters and receivers that could handle light at a wavelength of 1.3 microns. This change reduced attenuation because silica glass is more transparent at 1.3 microns than at 0.87 micron.

In 1982 the third generation began to appear as researchers developed manufacturing techniques that increased the purity of silica glasses. By doing so, scientists greatly enhanced the transparency of fibers in the wavelength region from 1.2 to 1.6 microns. To take advantage of this advance, they perfected the technology of the laser diode chip—a high-speed, monochromatic source of laser light. In particular, laser diodes could generate light of 1.55 microns, the wavelength at which silica glass is most transparent. These improvements raised the transmission capacity to hundreds of gigabit-kilometers per second.

By 1984, many researchers believed the basic design of lightwave communications systems had been worked out. But to their surprise, the fourth and fifth generations emerged from a community of scientists working on forward-looking approaches and novel devices.

The fourth generation consists of innovative systems that modulate the frequency or phase of light instead of its intensity. Because such systems preserve the phase and frequency of light, they are described as coherent, just as lightwaves that travel together with the same phase and frequency are called coherent. Systems based on intensity modulation are known as direct detection, because the detectors used in these systems can translate a change in intensity directly into a fluctuation in electric current.

In the laboratory, coherent systems outperform direct-detection systems. Coherent receivers can accurately detect light signals whose intensity is much lower than the minimum that direct-detection systems can discern. Signals can therefore travel farther through a coherent system than through a comparable direct-detection system.

By 1987, coherent systems had transmission capacities almost as high as those of direct-detection schemes. It was far from clear, however, whether either scheme could be improved. Some investigators favored coherent systems because of their greater potential. But they faced serious technical obstacles. They needed to produce light sources whose frequency and phase were very stable. And they had to develop devices that could modulate and detect the phase and frequency of light as easily as conventional components could change and monitor the intensity of radiation. Indeed, researchers had found devices that could accomplish these tasks, but at the time the approaches were too complex and expensive for practical considerations.

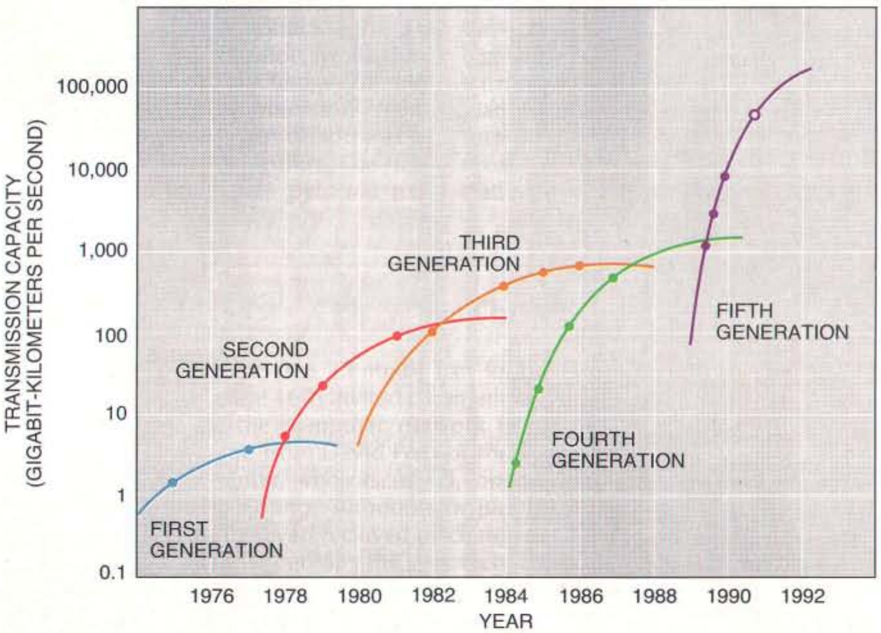
Workers were also struggling to improve direct-detection systems. They knew that if they could invent powerful amplifiers, they could compensate for the limited sensitivity of direct-detection systems. The amplifiers would boost the signals, enabling them to travel over increased transmission lengths.

Many researchers were attempting to build an optical amplifier on a single chip. They succeeded but had difficulty making the devices perform on the lab bench. It was hard to imagine that optical amplifier chips would be implemented on a large scale within several years.

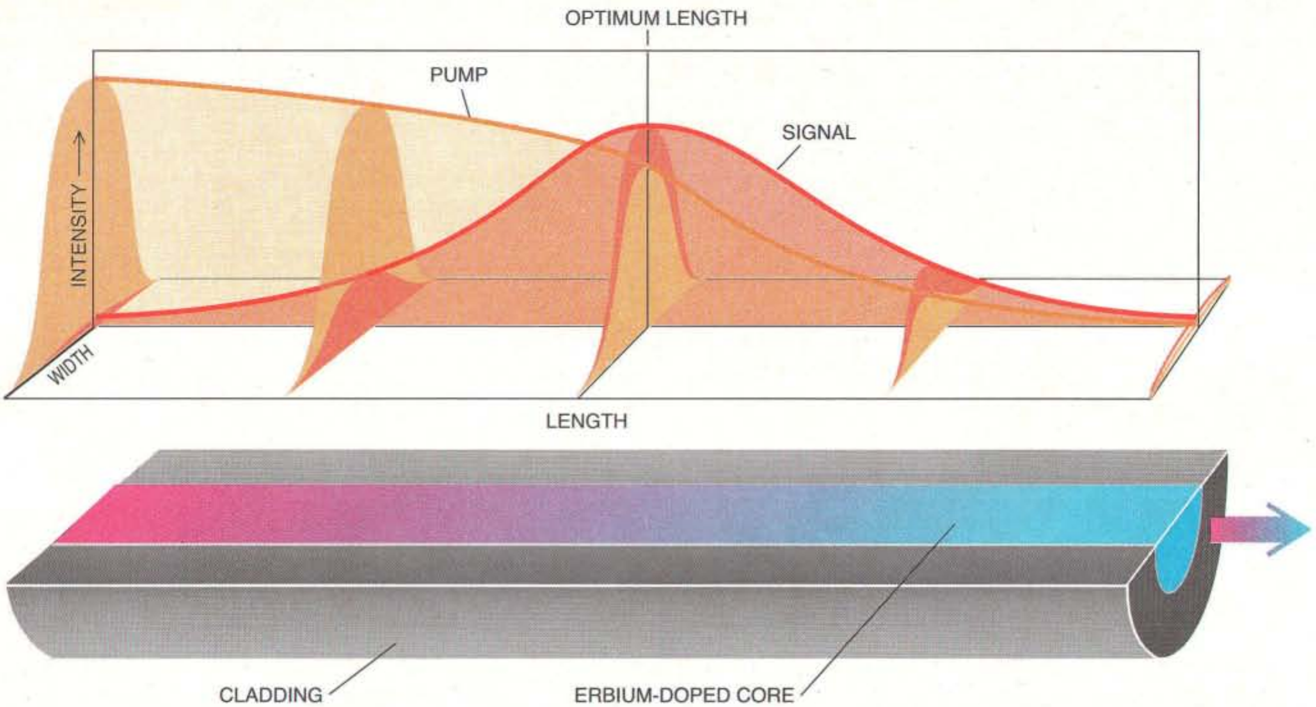
At the end of the 1980s, it seemed likely that both direct-detection and coherent systems would benefit from optical amplifiers. Yet it was not clear what kind of amplifier would prove beneficial. Some researchers, therefore, turned their attention to "exotic" devices.

The breakthrough was the development of the erbium-doped fiber amplifier. The main components of the amplifier are laser diodes and lengths of erbium-doped fiber. The laser diodes power the amplifier by providing infrared radiation at a wavelength of 1.48 or 0.98 micron. The light is absorbed by the erbium atoms, "pumping" them to a high-energy level. When a weakened signal enters the erbium-doped fiber, the "excited" erbium atoms transfer their energy to the signal through a process that Einstein called stimulated emission. In this way, the erbium-doped fiber regenerates the signal.

Erbium-doped fibers are not considered to be "repeaters" even though they



TRANSMISSION CAPACITY of optical fibers has increased 10-fold every four years since 1975. The five generations of the technology are shown. The open circle (*purple*) represents the results from an experiment that simulated long-distance transmissions using a loop of fiber amplifiers and optical fibers.



FIBER AMPLIFIER requires a source of pumping radiation and a single-mode optical fiber whose core is doped with erbium. The fiber amplifies signals as erbium atoms absorb the pumping radiation and transfer their energy to the signals. Hence, the intensity of the pumping radiation decreases over the length of the fiber. As the signal travels through the fiber, its intensity increases until the intensity of the pumping radiation drops below a certain threshold.

act as signal regenerators. Repeaters convert light into electric current, amplify the current and transform it back into light. Erbium-doped optical fibers do not interrupt the path of the light signal as it propagates from transmitter to receiver.

The introduction of the erbium-doped fiber ushered in the fifth generation of lightwave communications systems. Transmission capacities shot up from hundreds to tens of thousands of gigabit-kilometers per second.

The idea of doping optical fibers with erbium and other rare-earth elements dates to the early 1960s, when Charles J. Koester and Elias Snitzer of the American Optical Company investigated fibers doped with the rare earth neodymium. They discovered that the fibers could be used as a lasing medium and could amplify light. To measure the amplification, they coiled a one-meter length of the fiber around a flash lamp. They found that infrared radiation at a wavelength of 1.06 microns could be amplified by a factor of 50,000, a gain of 47 decibels. (Amplification is usually expressed in decibels, which are defined as 10 times the common logarithm of the ratio of the output power to the input.) A decade later Julian Stone and Charles A. Burrus, Jr., of Bell Laboratories were the first to demonstrate that neodymium-doped fibers could be energized by laser diodes rather than by cumbersome flash lamps or water-cooled lasers.

But during the 1970s, this promising line of research was abandoned. The neodymium-doped fiber amplifiers did not meet the needs of the time. Whereas the fiber amplifiers operated in the range of 1.06 microns, researchers were concentrating on applications that would work at wavelengths close to 1.3 or 1.55 microns, the wavelengths at which silica is most transparent. Could fibers be doped with other elements to make a device that would operate in the proper wavelength range?

The issue was not studied until 1985–86, when David Payne and his coworkers at the University of Southampton revitalized interest in fibers doped with rare-earth elements. In particular, they demonstrated that an erbium-doped fiber could make a novel type of optical amplifier at 1.55 microns. One year later they achieved a 25-decibel gain at this wavelength by pumping an erbium-doped fiber with 60 milliwatts of red light at 0.65 micron.

Most scientists who learned about the work at Southampton probably considered it an interesting piece of physics but nothing very practical. They were justified in their criticism because it is difficult to generate intense red light at 0.65 micron; one must use a large, ponderous laser. But these critics overlooked the fact that erbium can be pumped at other wavelengths. In particular, erbium can be energized with infrared radiation, which can be produced by a very practical device, the laser diode chip.

I was very interested in the work of the Southampton team, having spent a few years experimenting with fiber amplifiers that boosted signals through a process known as Raman scattering. When I joined AT&T in 1986, I wanted to investigate whether the technology of fiber amplifiers could be refined to improve lightwave networks. Just before my arrival, John B. MacChesney and Jay Simpson, who were working in the laboratory, had patented a process for fabricating optical fibers made up of many constituents—including rare-earth elements. I collaborated with them and Philippe Becker, also at Bell Labs, to study erbium-doped fibers.

We used an argon-ion laser to pump an erbium-doped fiber with green light at 0.51 micron. By the middle of 1987, we obtained an encouraging 22-decibel gain using 100 milliwatts of green light. But such results did not generate excitement among network engineers, who justifiably considered the device impractical.

My colleague Randy Giles and I then made a rather unorthodox decision. The next logical step would have been to give up the approach or to work toward building a more practical fiber amplifier. Instead we wanted to see how well the amplifier could transmit data. We proceeded to measure the error rate of data flow through the strange, green-glowing fiber amplifier. After spending a few hours fighting spurious effects, we obtained an error rate of one part in a billion at a data flow of two gigabits per second. This rate is the industry standard for "error-free" communications.

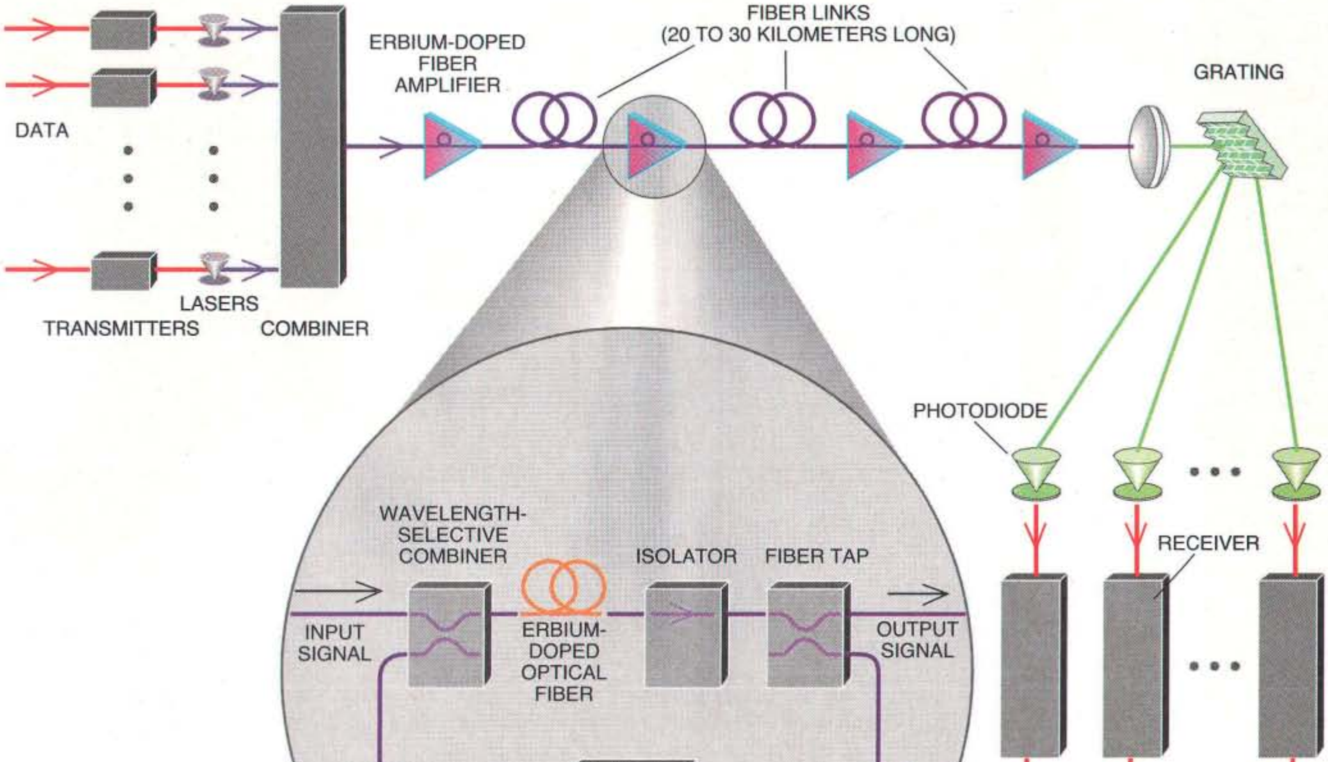
Most important, the observed gain did not depend on the polarization of the signal, whereas the semiconductor amplifiers that have been developed are sensitive to polarization. This result was the first sign that erbium-doped fibers might have a real advantage over other amplifiers.

Encouraged by these findings, we conducted an experiment to determine whether erbium-doped fibers could regenerate several signals at the same time. We found that they could. The signals did not interfere with one another, or more specifically, the cross talk between signals was negligible. We had identified a second advantage of erbium-doped fiber amplifiers.

During the late 1980s, my colleagues and I set out to construct a practical erbium-doped fiber amplifier. First and foremost, we need to prove that compact laser diodes could replace massive lasers as an efficient means for powering erbium-doped fiber amplifiers.

Our efforts were complicated by the fact that although several different wavelengths of radiation can be used to energize erbium, only a few do so efficiently. For instance, when an erbium atom is pumped with light of 0.67 or 0.80 micron, it absorbs a photon, achieving the desired energy state, but it may then jump to a second state by absorbing a second photon. In a fiber amplifier, a fraction of the erbium atoms will jump to this second state, which prevents them from transferring their energy to the optical signal. As a result, the efficiency of the amplifier is reduced.

Luckily, by a trick of nature, radiation at 0.98 or 1.48 microns can energize erbium without suffering from this unwanted effect. These wavelengths can be produced by semiconductor laser diodes fabricated from compounds of indium, gallium, arsenic and phosphorus. It is also fortunate that a silica fiber can be designed so that light at 0.98, 1.48 or 1.55 microns will travel in a single mode. This type of propagation is desirable because it allows the fiber to carry high-intensity light, which can energize the erbium atoms with great efficiency.



LONG-DISTANCE COMMUNICATIONS SYSTEMS will rely on erbium-doped fiber amplifiers in the near future. In such systems, several optical transmitters generate signals, which are combined and directed into a single fiber. Fiber amplifiers are placed between ordinary strands of optical fiber and boost the signals periodically. When the signal emerges from the array of fibers, it is focused onto an optical grating, where it is separated. The split signals are then guided toward several receivers. In the inset, which depicts the details of the fiber amplifier, a photodiode measures the intensity of the output signal and provides feedback to keep the amplification at a constant level.

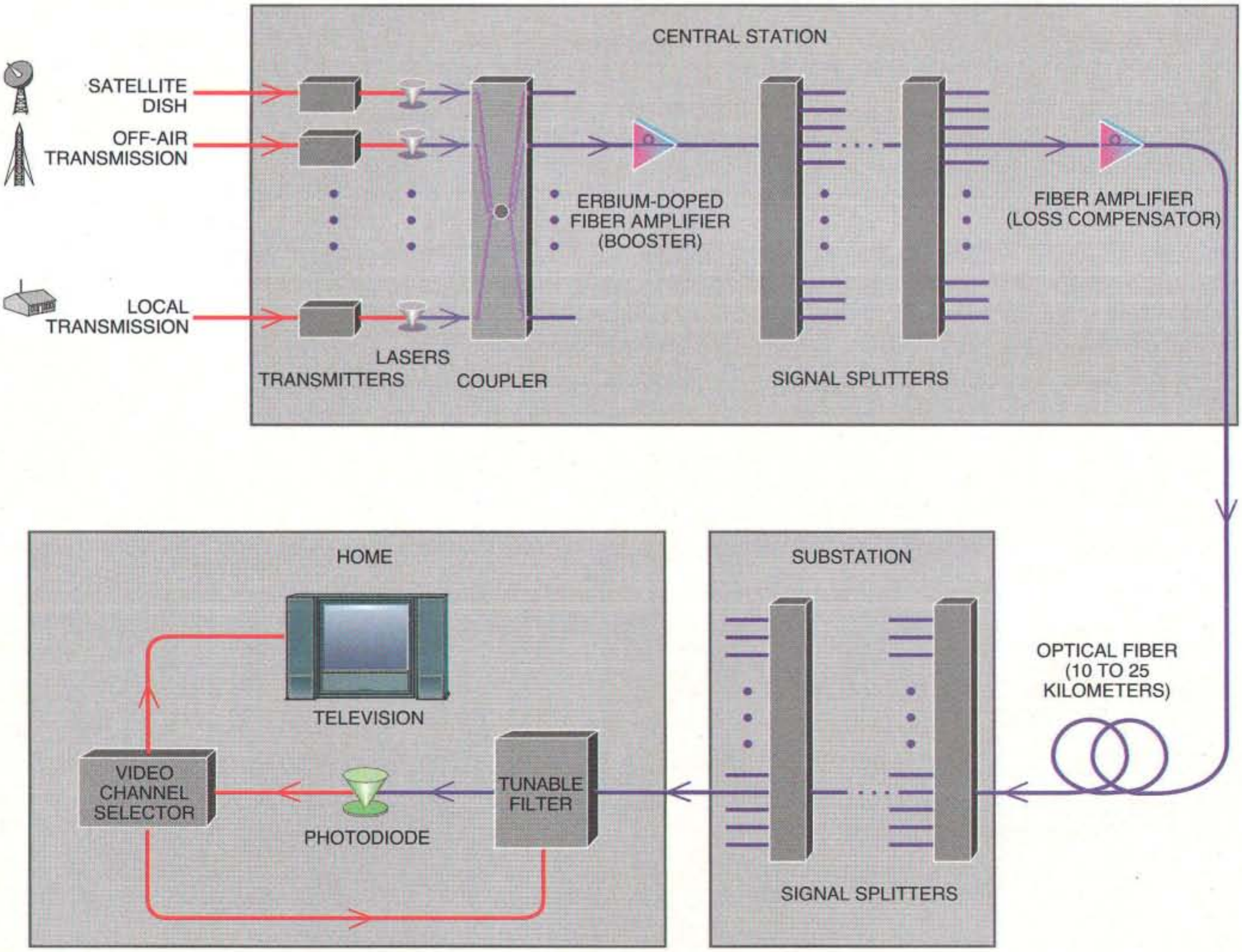
To determine whether laser diodes might be practical, my group and others first used conventional lasers to test how erbium-doped fibers would perform powered by radiation of either 0.98 or 1.48 microns. Researchers at Southampton conducted many of the first experiments at 0.98 micron, which they identified as the most efficient wavelength for pumping. Meanwhile my co-workers and I focused on 1.48 microns, the wavelength for which the laser diode technology was more mature. Other institutions, including the Nippon Telegraph and Telephone (NTT) Research Laboratories near Tokyo, joined the investigation. So far the best results, as reported by AT&T for the 1.48-micron wavelength and NTT for the 0.98-micron wavelength, are that an erbium-doped fiber amplifier powered by a mere 10 milliwatts of power can achieve a gain in the range of 30 to 40 decibels. Such power levels were shown to be well within the capability of laser diodes, which can now generate more than 200 milliwatts of light at 0.98 and 1.48 microns.

Perhaps most significant, in 1989 NTT was the first to demonstrate that an amplifier could be efficiently pumped with a laser diode. The same company successfully tested a prototype system of erbium-doped fiber amplifiers and ordinary fiber links. The system carried about two gigabits per second over a distance of 212 kilometers, at the time.

Once NTT had shown that the transmission capacity could be greatly increased, research groups began competing for world records as fiercely as Olympic teams. In the middle of 1989 both NTT and Bell Communications Research reported transmission rates of about 10 gigabits per second over distances of 150 kilometers, bringing the maximum transmission capacity to 1,500 gigabit-kilometers per second. Near the end of 1989, workers at KDD announced even more startling news. By using a series of 12 fiber amplifiers, they were able to transmit data at 1.2 gigabits per second over the incredible distance of 904 kilometers.

Then, early in 1990, workers at NTT made a remarkable comeback. They were the first to construct a coherent system that included fiber amplifiers. The system could transmit data at 2.5 gigabits per second over 2,223 kilometers. The industry had entered the era of 1,000-kilometer-long systems.

But even these results do not demonstrate the full potential of erbium-doped fiber amplifiers. Using an experimental system that emulates a very long transmission link, Neal S. Bergano and his colleagues at AT&T achieved transmission capacities as high as



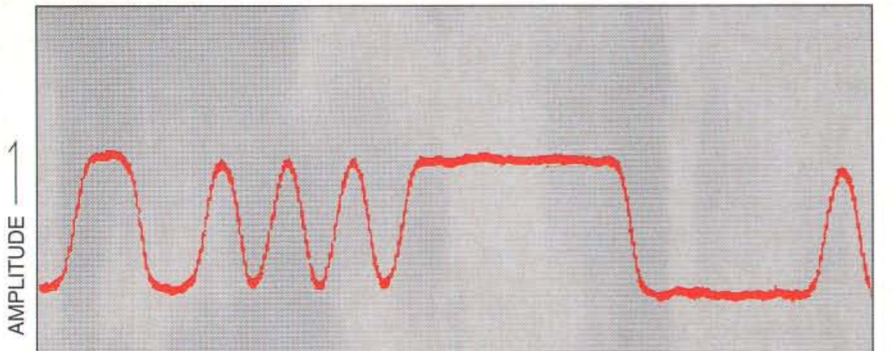
BROADCAST NETWORK for high-definition television and other communications services is one system that could benefit from erbium-doped fiber amplifiers. According to recent estimates, the system could transmit signals as far as 25 kilometers and could reach as many as 10 million users. Once the signals are received or generated, they travel to a series of transmitters and lasers that convert the signals into light. The signals are combined into one and guided into several ordinary fibers. Amplifiers boost the signal as it is split numerous times and as it decays over lengths of ordinary fiber. The signal ultimately reaches the user, who can select the desired portion of the signal.

50,000 gigabit-kilometers per second. This capacity is 100 times greater than what can be achieved so far in an optical fiber system without amplifiers, and it is 10 times greater than what has been attained in systems deployed across the oceans.

The group placed a few erbium fiber amplifiers between strands of low-loss fiber and joined the ends to form a loop. When they injected light pulses into the loop, they succeeded in transmitting 2.4 gigabits per second over 21,000 kilometers and five gigabits per second over 9,000 kilometers. The output signals emerged undistorted except for a small, unavoidable amount of noise generated by the fiber amplifier. Although these results were achieved using a fiber loop, which differs somewhat from a demonstration of point-to-point communications, the achievement indicates that fiber amplifier systems have enormous potential.

Whereas some investigators have focused their efforts on achieving record-breaking transmission capacities during the past four years, others have made significant progress in the study of solitons and their use in communications systems.

A soliton can travel indefinitely through a "perfect" fiber, that is, one that does not attenuate light. A soliton, like an ordinary light pulse, consists of many components that differ slightly in wavelength. A soliton also varies in intensity over its length, width and height. Lightwaves of different wavelengths travel through fibers at different velocities. This effect is known as dispersion. And when the intensity of light exceeds a certain level, lightwaves of the same wavelength but different intensities propagate at different speeds. This phenomenon is known as the optical Kerr effect. The principle behind solitons is that the optical Kerr effect can be used to counteract dispersion. If each wavelength component is generated so that it has a particular intensity, all the components can be made to travel at the same speed. As a result, the soliton will neither broaden nor compress as it travels through a lossless fiber.



SIGNALS (*top*) traveled 9,000 kilometers through a loop of ordinary fiber and erbium-doped fiber amplifiers, emerging (*bottom*) virtually undistorted and free from noise. These results, which were obtained by Neal S. Bergano and his co-workers at AT&T Bell Laboratories, show the great potential of fiber amplifiers.

Of course, all fibers attenuate signals to some degree, and the soliton will slowly decay as the pulse propagates. Yet it will retain its shape until the power of the soliton drops below a certain level.

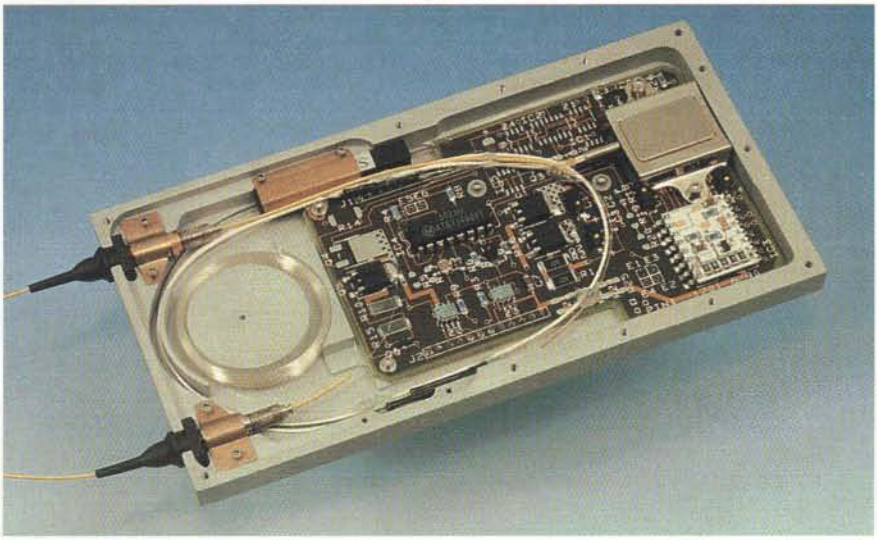
Linn F. Mollenauer, who pioneered the field of solitons in optical fibers, proposed that fiber amplifiers could compensate for the losses in fibers, making it possible to sustain the propagation of solitons over long distances. In 1988 he proved it at AT&T. By trapping solitons in a 40-kilometer loop of fiber and fiber amplifiers, Mollenauer and his colleagues ultimately observed solitons that propagated for 10,000 kilometers with little broadening. Recently they showed that solitons enable information to be transmitted, error free, at a rate of 2.5 gigabits per second over a distance of 14,000 kilometers. This result gives every indication that solitons could be used for long-distance communications.

Solitons are the ideal light pulse for communications. They can be spaced closely together because they do not merge into one another and because they are short (about 10 trillionths of a second). Soliton systems—in which the presence or absence of a soliton represents a single bit of information—should attain transmission capacities of at least five gigabits per second over distances of 10,000 kilometers.

Soliton systems should achieve even greater capacities if they can take advantage of other emerging technologies. Signals transmitted as solitons enable several channels of information to be sent simultaneously at different wavelengths. This approach, called wavelength-division multiplexing, is impractical with nonsoliton signals because the signals disperse and mix irreversibly with one another over long distances. In soliton systems the maximum number of channels is ultimately limited because fiber amplifiers boost signals of different wavelengths by different amounts. A soliton system could realistically support at least five channels and could therefore carry 25 gigabits per second.

Mollenauer and his colleagues have also investigated "polarization multiplexing" for soliton systems. In such schemes the number of channels can be doubled by transmitting signals that have the same wavelength but two different polarizations. This approach would bring the total transmission rate to the high rate of 50 gigabits per second. The techniques used for generating, modulating and detecting soliton pulses are still being developed in the laboratory. For this reason alone, soliton systems may not be deployed for some time.

Another alternative for the future of communications is coherent systems. Coherent techniques could exploit the entire wavelength region from 1.2 to 1.6 microns and would allow optical channels to be packed very closely together. To make a practical system, workers must still develop frequency-agile devices, which have better absolute frequency control and stability. If progress continues in this field, workers may well find a way to eliminate the need for optical amplifiers in local-area networks. It seems clear, however, that long-distance communications will continue to depend on optical amplifiers. Furthermore, it is too early to predict when frequency-agile devices will be ready for use.



FIBER AMPLIFIER can boost telephone signals through fiber-optic cables that span continents. The amplifier contains an erbium-doped optical fiber, which is wound around a spool at the left. The fiber is energized by light from a laser chip, which is housed in the small metallic case at the top right. Light signals enter and exit the fiber through the yellow cables at the left. The device, which is 15 centimeters long, was designed by Terry W. Cline and his colleagues at AT&T Bell Laboratories.

Erbium-doped fiber amplifiers are already having a tremendous impact on long-distance communications. And in the next decade fiber amplifiers should play an even greater role in the field of lightwave networks. Indeed, the technological challenge of such networks is to exchange the maximum amount of data among the largest possible number of users in the shortest interval of time with the smallest probability of failure. For this type of communication, the advantage of employing optical fibers, rather than microwave cables, comes from their low loss and their ability to convey data at higher bit rates over longer distances.

The simplest form of a fiber network is called a broadcast network, in which optical data are transmitted from a central station to an ensemble of users. A fiber amplifier could compensate for losses as the signals from the central station are split over and over again. Other fiber amplifiers in the branches of the network could compensate for propagation losses over extended distances of 10 to 25 kilometers, permitting the scale of the network to be increased. The number of homes that could be reached through such a network is enormous.

The company British Telecom has recently demonstrated the broadcasting of 384 video channels to nearly 40 million potential users within a 50-kilometer-diameter area. Such systems can be expanded in area by increasing the fiber length and using more fiber amplifiers to compensate for the loss. The maximum number of users is limited, however, because as the signal passes through one amplifier after another, it eventually accumulates so much noise that it becomes unusable for communications purposes.

Networks that include fiber amplifiers can transmit data at very high rates because they are able to support many channels. The number of channels is not reduced by dispersion, which is negligible over the short distances covered by fiber networks. The number of channels is finite, however, because of four effects. First, if the combined strength of the signals from each channel exceeds a certain level, the erbium amplifier will not amplify the signal. Second, the fiber amplifier can transmit only a limited range of wavelengths (from 1.53 to 1.55 microns). Third, the wavelengths of any two optical channels must differ enough so that the channels do not interfere or mix together. And finally, the gain of the amplifier varies with wavelength. The number of channels may also be reduced because of limitations in the ability of photonic devices to discriminate optical channels.

Despite these limitations, lightwave networks that employ fiber amplifiers may ultimately carry an estimated 2,000 to 3,000 gigabits per second. So far no other technology can compete with these rates.

As erbium fiber amplifiers and other lightwave technologies are introduced into long-distance systems and local-area networks during the coming decades, people will have access to new types of powerful communications services. Lightwave communications systems will integrate the functions of telephones, facsimile machines, computers and televisions. They will also offer many new services: Businesses will have access to vast data banks and will be able to take advantage of interactive video networks for teleconferencing. Researchers will be able to exploit fully the resources of supercomputers located thousands of kilometers away. Homeowners will be able to choose programs from video libraries. It is reasonable to infer that just as telephones and televisions have transformed work and leisure in this century, lightwave communications systems promise to influence society profoundly in the 21st century.

FURTHER READING

BIBLIOGRAPHY ON ERBIUM-DOPED FIBER AMPLIFIERS: 1987 TO 1990. Harish R. D. Sunak. EDFA Consultants (P.O. Box 1668, Kingston, R.I. 02881-1668), 1991.

FUNDAMENTALS OF PHOTONICS. Bahaa E. A. Saleh and Malvin Carl Teich. John Wiley and Sons, 1991.

MULTIGIGABIT SOLITON TRANSMISSIONS TRAVERSE ULTRALONG DISTANCES. L. F. Mollenauer, J. P. Gordon and S. G. Evangelides in *Laser Focus World*, Vol. 27, No. 11, pages 159–170; November 1991.

RARE EARTH DOPED FIBERS AND DEVICES. Edited by M. J. Digonnet. Marcel Dekker, 1993.

Microlasers

Millions of lasers measuring just a few millionths of a meter can now be etched on a single chip, offering a host of novel applications in optical communications and information processing

by Jack L. Jewell, James P. Harbison and Axel Scherer

Investigators have long dreamed of building an optical computer. Electronic switches and circuits would be replaced by a network of light through which individual impulses carry and process information. The idea is not only aesthetically appealing but also offers the promise of a machine that would be even faster and more versatile than the most powerful computer in use today.

The realization of the dream, however, has been thwarted by the lack of sufficiently tiny optical-signal processing devices. Like its electronic counterpart, the transistor, the basic building block of any optical processing or communications system must operate at low power and at high speed, and to do so it must be small. In the microelectronics industry, transistors having dimensions smaller than a micron (a millionth of a meter) are now routinely fabricated in numbers approaching tens of millions on a single semiconductor chip. Compared with this astonishing feat, the miniaturization of optical devices has lagged considerably.

The current state of affairs in commercially available optical technology is embodied by a device called the semiconductor diode laser, which is found in everything from compact-disc players to fiber-optic communications systems. Although the diode laser has revolutionized the storage and communication of information, the size of the device and its degree of integration are roughly comparable to that of the individually packaged transistors found in a late 1950s radio. A typical diode laser measures a few microns wide by several hundred microns long—making it several hundred times bigger than one of its microelectronic counterparts. Although such a device is orders of magnitude smaller than the familiar red helium-neon laser commonly used in supermarket bar-code scanners, it is simply too big to be useful in an optical computer.

Very recently, significant advances have been made in the miniaturization of diode lasers. In May 1989, in work that evolved from efforts to construct two-dimensional arrays of optical switches in hopes of eventually building optical computers, we fabricated more than one million microlasers, or micron-size lasers, on a single semiconductor chip about seven millimeters wide by eight millimeters long [*see illustration on opposite page*]. The microlasers were proposed by one of us (Jewell) and Sam McCall of AT&T Bell Laboratories. They were created at Bell Communications Research (Bellcore) by the other two of us and Leigh Florez. The devices range in size from one to five microns.

Such a scale is already two orders of magnitude smaller than that of conventional diode lasers. With additional work, the size can probably be brought down by another order of magnitude. Perhaps most exciting, as the microlaser approaches its practical lower limit in size (probably between one half and one quarter of a micron), it is believed that the quantum mechanical process of light emission can be radically altered in a way that will further enhance the performance of the device.

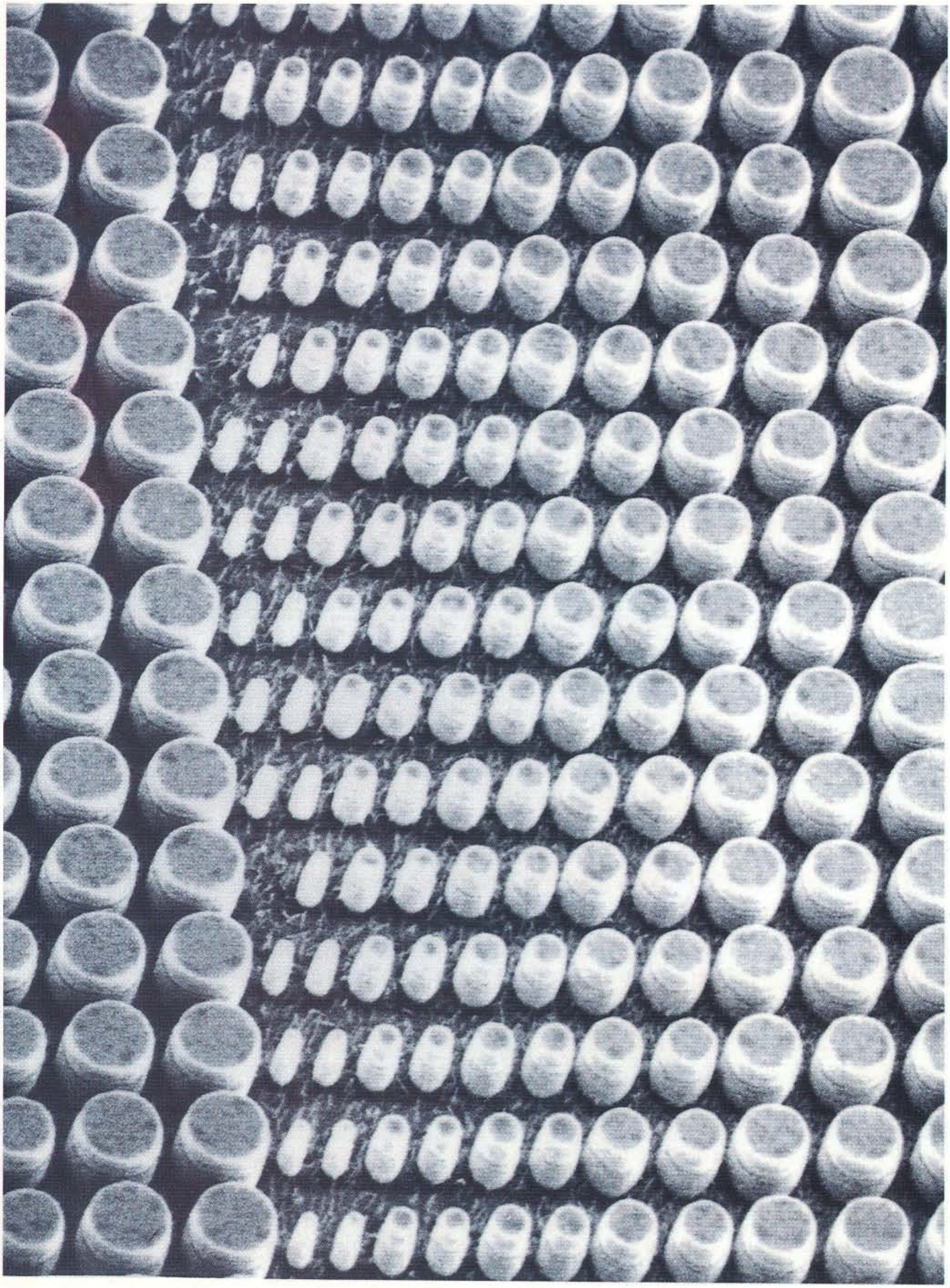
Microlasers are so new that at this time no one can predict whether they will make a significant impact on the commercial market. Moreover, the development of an optical computer is far from an easy task [see "The Optical Computer," by Eitan Abraham, Colin T. Seaton and S. Desmond Smith; SCIENTIFIC AMERICAN, February 1983]. Nevertheless, microlasers have generated enormous excitement for their potential applications in optical computers as well as in optical communications and information processing in general.

The principles of operation underlying a diode laser are the same as those for any laser [*see top of illustration on page 64*]. Atoms in a part of the laser called the amplifying medium—typically a solid, liquid or gas—are pumped, or energized, either electrically or with a source of electromagnetic radiation. When a light wave of a specific wavelength traveling through the amplifying medium encounters a pumped atom, it can induce the atom to release its energy in the form of a light wave at the same wavelength. The process is coherent, which is to say that the crests and troughs of the waves match up, and the intensity of the light increases. Mirrors on each end of the amplifying medium form a cavity, and they force the light to bounce back and forth many times through the medium, maximizing the increase in intensity.

JACK L. JEWELL, JAMES P. HARBISON and AXEL SCHERER share an interest in optoelectronics. Jewell received a Ph.D. in optical sciences from the University of Arizona in 1984. Before co-founding Photonics Research, Inc., he was at AT&T Bell Laboratories. Harbison holds A.B., S.M. and Ph.D. degrees from Harvard University. He went to Bell Laboratories in 1977 and joined Bellcore at its inception in 1984. Scherer is associate professor of electrical engineering at the California Institute of Technology. He earned a Ph.D. from the New Mexico Institute of Mining and Technology in 1985.

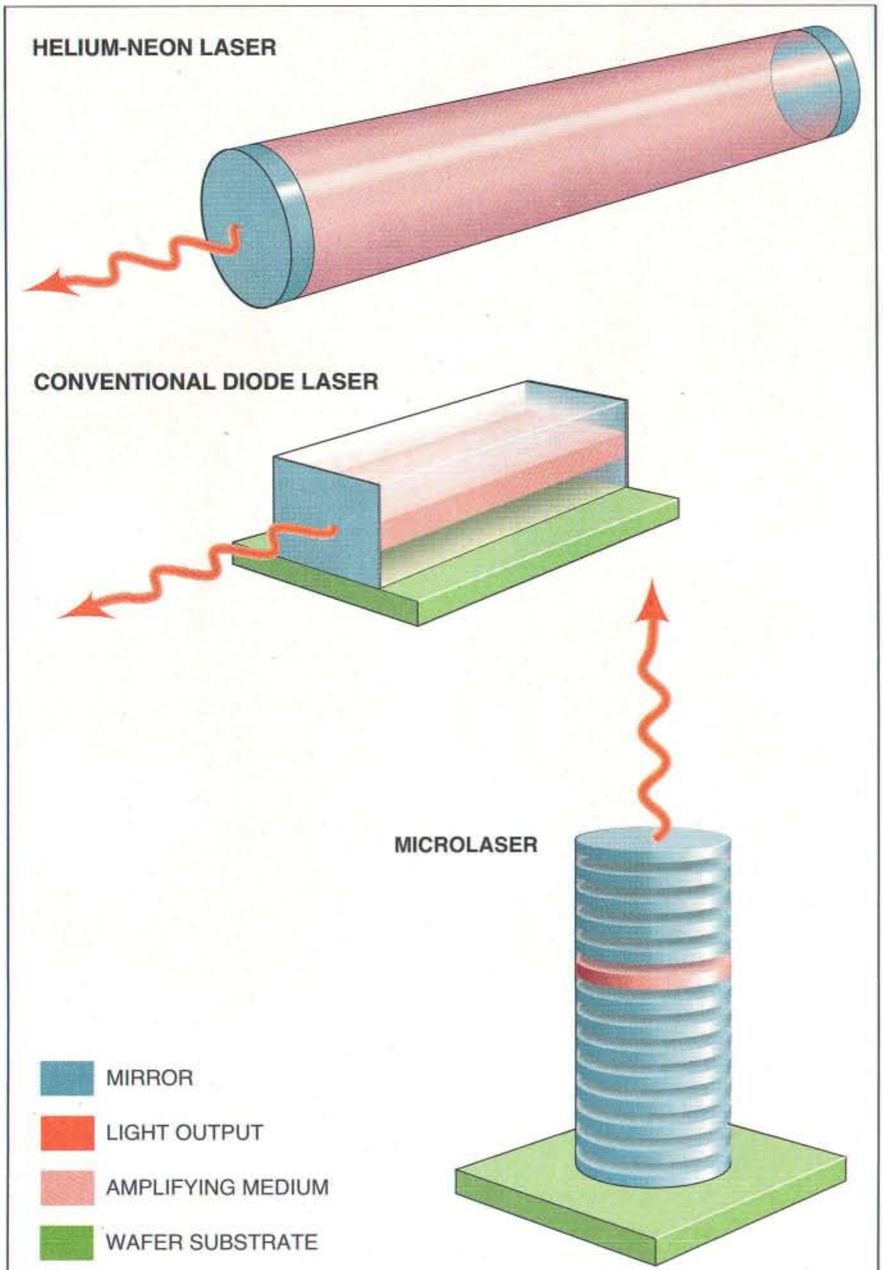
FIRST ARRAY OF MICROLASERS was completed in May 1989 by the authors. A tiny portion of the array, which contained more than one million microlasers, is shown in this scanning electron micrograph. The largest devices have a diameter of about five microns (millionths of a meter); the smallest are about a micron. Laser beams, which emerge from the circular-shaped ends, were generated in all the devices except the very smallest.

Reprinted from the November 1991 issue



The reflectivity of one or both of the mirrors is less than unity, with the result that some of the intensified light is able to escape from the cavity in the form of a laser beam.

In a diode laser the amplifying medium is a long, brick-shaped diode, a device that permits electric current to flow freely in one direction but blocks the flow of current in the other. The diode is made from a semiconducting material such as gallium arsenide. The amplifying medium of a conventional diode laser is about as long as the laser itself [*see middle of illustration on this page*]. Mirrors are formed on each end of the "brick" by simply cleaving the wafer: each interface between the semiconductor and air has a 30 percent reflectivity, which is more than sufficient for operation. When an electric current is made to flow parallel to the mirrors, the diode fills with electrons and holes (parts of the semiconductor lacking electrons). A short time later the electrons and holes recombine, and when they do, they emit light. The light is intensified by multiple reflections through the semiconductor, and a portion of it emerges from one of the mirrors in the form of a laser beam parallel to the underlying wafer substrate; that is, the beam comes out of the edge.



PRINCIPLES OF OPERATION are the same for a conventional helium-neon laser found in a supermarket bar-code scanner (*top*), a conventional semiconductor diode laser in a compact-disc player (*middle*) and one of the newly developed microlasers (*bottom*). Atoms in an amplifying medium (*pink*) are pumped, or energized, by electromagnetic radiation or current. When a light wave traveling through the medium encounters a pumped atom, it can induce the atom to release its energy in the form of a light wave at the same wavelength. The intensity of the light increases as it bounces back and forth between two mirrors (*blue*) positioned on each end of the medium. The reflectivity of one of the mirrors is less than unity, which allows some of the intensified light to escape in the form of a laser beam (*red*). The lasers are not drawn to scale: a helium-neon laser is about 100 to 1,000 times longer than a conventional diode laser, which in turn is about 100 times longer than a microlaser.

One characteristic that makes a microlaser so very different from a conventional diode laser is that the laser beam emerges perpendicular, instead of parallel, to the wafer substrate [*see bottom of illustration at left*]. Instead of the brick-shaped appearance of a conventional diode laser, a microlaser looks like a tiny Coke can; the laser beam shoots out of the top (or out of the bottom through the transparent wafer on which it has been grown). The difference in geometric construction means that more lasers per area can be packed on a wafer. Kenichi Iga and his colleagues at the Tokyo Institute of Technology first demonstrated the feasibility of such a surface-emitting (instead of edge-emitting) approach in the late 1970s.

Our own work in microlasers dates to about the time of Iga's breakthrough. Investigators at Bell Labs built surface-emitting lasers by sandwiching films of gallium arsenide that measured from one eighth of a micron to a few microns thick between mirrors and gluing the entire stack together. To sensible people, the notion of turning such a procedure into a practical technology was silly; fabrication bordered on black magic, and the devices were of poor quality. (Of course, the first transistors also had these "attributes.")

Steady improvements turned the tide. A microlaser, including mirrors, is now typically about six microns long, compared with the 250-micron length of a conventional diode laser. But what actually makes a microlaser "micro" is that its Coke-can geometry serves as a waveguide: an optical "hose" that prevents light from spreading outward as it propagates. (An optical fiber is a well-known example of a waveguide.) The typical diameter of a microlaser, as we mentioned earlier, is a few microns.

Even though microlasers are still in

their infancy, their small size makes their power requirements comparable to those of the more mature conventional diode lasers. An important yardstick used to characterize the power requirements of any diode laser is the threshold current, the flow of electric current needed to make the diode just barely lase. The lowest threshold of a microlaser is 0.7 milliampere, compared with the 0.5-milliampere threshold of a conventional diode laser. With additional work, it is hoped that the threshold of a microlaser, because of its exceptionally small size, can be lowered to a few millionths of an ampere. At present, microlasers also fall somewhat short of conventional diode lasers in their bit rates—the number of zeros and ones that can be generated per second. At a bit rate of five billion per second, microlasers are not quite as fast as the best diode lasers, but rates approaching 100 billion per second are expected to be achieved eventually.

How are such exquisitely small, high-performance lasers made? Two crucial techniques make it possible. The first technique, molecular-beam epitaxy, allows the basic material of each laser to be built up from layers of semiconducting materials added one at a time. Individual lasers numbering in the thousands or millions can then be stamped out simultaneously by means of deep vertical etching, in conjunction with conventional photolithographic patterning.

Conceptually, molecular-beam epitaxy is quite straightforward. A semiconductor crystal is placed in a vacuum. Containers are filled with each of the chemical elements, such as indium, gallium, aluminum and arsenic, that will be used to build the desired semiconductor layers on top of the substrate. When an individual container is heated to the prop-

Fabricating and Testing the First Microlaser

Author Jack Jewell tells how it felt to do the groundbreaking work.

After months of strategizing and designing—and mostly anticipating—Axel Scherer picked up the laser wafer from Leigh Florez, who—working with Jim Harbison—had grown the structure by molecular-beam epitaxy. The wafer was two inches in diameter but only one fiftieth of an inch thick. We practically tiptoed with it back to Axel's lab. In spite of our care, however, within five minutes we managed to shatter the fragile disk into many odd-size pieces. Axel and I looked at each other and could think of nothing else to do but laugh. After all, we had planned to carefully cleave the wafer into smaller pieces anyway; our bad luck saved us time.

We picked up the larger pieces and started processing them. By midnight, we were ready to etch individual microlasers on the fragments. Then misfortune intervened in our favor once again. Axel told me that an important part of the etching apparatus, called a neutralizer, was not working and could not be fixed. Because the neutralizer sounds like some kind of gadget out of "Star Trek," we decided to pretend that it was not important and go on without it. By 2 A.M., Axel had etched several of our lasers, and we took the scanning electron micrograph shown on page 63. In the past, ugly devices that had looked like micro-trash cans had often worked quite well, whereas nice-looking devices had not worked at all. We agreed that the ones we had just made didn't stand a chance. The final step in the fabrication was to apply electrical contact points to the tops of the devices. For that, we would have to wait for the next day, since we knew of only one person who could do the job, Ray Martin (another Bellcore worker).

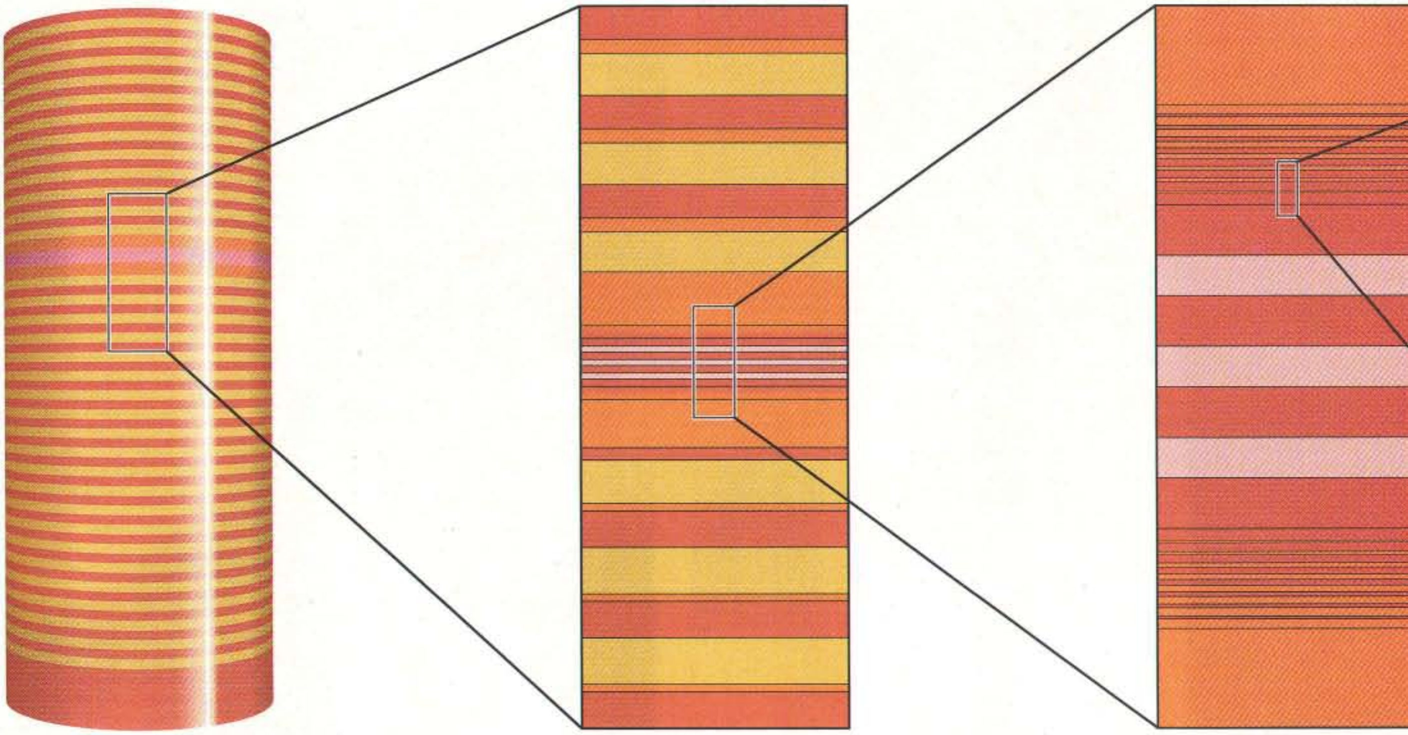
After a brief, fitful sleep, we asked Ray to spot-weld the contacts in place. At this point, anxiety really began to set in. We were rapidly approaching the transition from hopeful fabrication to the final test, which would surely spell failure and send us into months of troubleshooting. It didn't help that almost everyone who knew what we were trying to do thought we were more than a little daft. My wife (who was not one of those people) called to ask me if the lasers worked. I growled that we hadn't had time to test them yet and that they wouldn't work anyway, so don't bother me!

To test the lasers, we needed to touch the tops with a needlelike electrical probe. Trying to apply electric current through devices on my optical bench was a novel experience, because my entire career has been devoted to optics. Applying current to devices five microns tall and between one and five microns in diameter was definitely not for beginners. Axel and I took refuge in the fact that we were trying so many new things at once that no matter what happened we were sure to learn something. On touching the probe tip to one of the largest devices, we saw that current flowed in one direction, but not the other. Good! At least the device acted like a diode. We decided to set up a television camera to see if the desired infrared light was emitted. It was! The light was very dim up to about one or two milliamps of current and then increased dramatically, an encouraging sign. We also saw interference patterns, which are a characteristic of laser light.

Something must be terribly wrong. The gadgets were behaving exactly like lasers. This could not be possible. We tried another device, then another and another. They all acted more or less the same, all over the chip. This could not be possible. It was too easy. Things don't work like that. Yet there we were.

For the next few weeks, it was hard to do any serious work. We found that with the aid of a device called a translator we could run the microlasers under the electrical probe at a rate of about a dozen a second. With the turn of a single knob, the lasers would blink on and off like a string of light bulbs. Over and over again, we went back and forth on the same devices, but they showed no signs of wearing out. Every time we decided to make some quantitative measurements, our efforts would quickly degenerate into turning the translator knob and watching the blinking lights, still in disbelief.

"Our efforts would quickly degenerate into turning the translator knob and watching the blinking lights."



SUCCESSIVE MAGNIFICATIONS of a microlaser reveal the large number of semiconducting layers—typically 500 or more—from which it is fabricated. In the most highly magnified drawing (*right*), individual atoms are visible: arsenic (*gray*), aluminum (*yellow*) and gallium (*red*). Layers made from them appear in the other three drawings: aluminum arsenide (*yellow*) and gallium arsenide (*red*), plus the alloy mixtures of aluminum gallium arsenide (*orange*) and indium

er temperature, the element within it begins to evaporate, streaming out the open end of the container and into the vacuum in much the way that steam escapes out the hole of a heated teapot. The result is a wide beam of molecules that sprays in the direction of the wafer. When desired, each beam of molecules can be blocked by a mechanical shutter.

The formation of multiple layers of different semiconductor materials is then accomplished by simply determining which chemical elements should be in which layer and opening the appropriate shutters. For example, layers of aluminum arsenide are formed by opening the aluminum and arsenic shutters simultaneously while keeping all the other shutters closed; similarly, layers of gallium arsenide are formed by opening the gallium and arsenic shutters. The thickness of each layer is determined by how long the shutters are held open. Molecular-beam epitaxy is such a refined technique that it is possible to grow an individual layer having the thickness of a single atom.

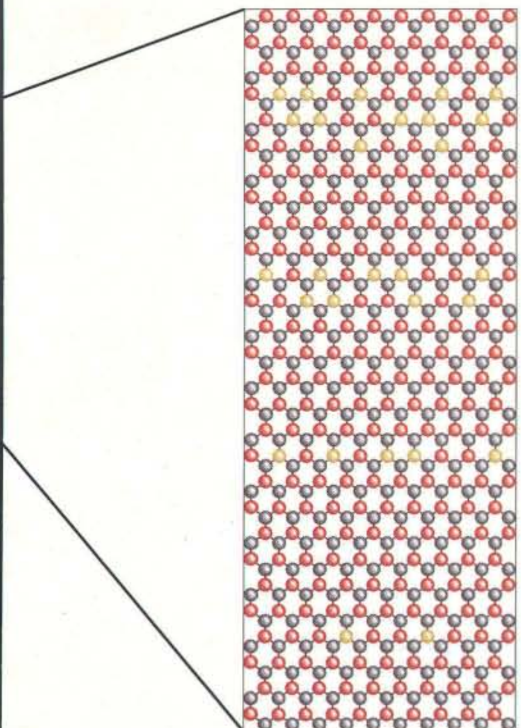
A typical microlaser may consist of more than 500 individual layers. To minimize the power required by the laser, one makes the amplifying medium quite thin, about a hundredth of a micron. The fact that it is so thin means that for the laser to work, light must bounce back and forth many more times than it must for a conventional diode laser. As a consequence, the mirrors must have a much higher reflectivity than the 30 percent reflectivity of the conventional diode laser. In fact, the reflectivity must be 99 percent or more.

Such a high reflectivity is attained by depositing in alternate layers two semiconductors, such as gallium arsenide and aluminum arsenide, that have different indices of refraction—that is, light travels at different speeds through them. The difference in the indices gives rise to a partial reflection at each interface, just as the difference in the indices of glass and air causes a partial reflection from a window. Although the reflection from each interface is only about 0.6 percent, a total reflectivity of greater than 99 percent can be achieved by stacking many alternate layers of the proper thickness.

Ironically, then, the bulk of a microlaser consists not of the approximately hundredth-micron-thick section containing the amplifying medium but rather of the mirrors, which together are four or five microns thick. The relative thickness of the mirrors is well justified, however, because an entire array of microlasers—mirrors and all—can be grown at once, instead of being painstakingly fabricated in the series of steps required to make a diode laser.

Just as important as the ability of molecular-beam epitaxy to form an entire microlaser in one step is the precision with which it can form the thickness of each layer. Careful control of layer thickness is particularly important in fabricating the mirrors. The technique for achieving the necessary control is based on the fact that an individual atomic layer becomes progressively rougher as it grows, until about half of it is in place. Then the layer begins to become progressively smoother until, when it is completely in place, it is atomically smooth.

One can determine how smooth a surface is by bouncing a beam of high-energy electrons off it: the smoother the surface, the more easily the electrons are reflected and detected on a screen on the other side. Hence, the intensity of the reflected beam oscillates in time as successive layers are grown. Accurate measurement of the frequency of the oscillation can precisely determine the time to deposit a single layer



gallium arsenide (*pink*). The actual amplifying medium is quite thin (*pink layer*); the bulk of the device consists of its mirrors (*red and yellow layers*).

of atoms. The resulting exact knowledge of the deposition rate can then be used to strictly control the layer's thickness. Timing of the mechanical shutters with a computer provides precise regulation.

Another advantage of molecular-beam epitaxy is that it enables one to modulate the electrical conductivity and reflectivity of the mirrors. The electric current for pumping the diode must pass through the mirrors, but unfortunately, the current does not flow easily in this instance. The electrical and optical properties cannot be optimized simultaneously, and so microlasers must be designed to accommodate the two competing demands as well as possible.

Once the wafers have been grown, conventional techniques for transferring patterns are used to carve the wafers into individual microlasers. An etch mask, consisting of a two-dimensional array of disks of etch-resistant material, is deposited on the wafer surface by vapor deposition and photolithography, a printing process that uses plates made according to a photographic image. Then, in a novel etching step, a collimated beam of xenon atoms, guided by the deposited mask, vertically etches out each Coke can-shaped microlaser. The resulting sidewalls are smooth, and the surface damage is minimal.

Our group at Bellcore has used these advanced techniques to make arrays of microlasers in which a few layers of one of the mirrors were deliberately varied in thickness. Consequently, the wavelength of each laser differs slightly from that of its neighbor—by an evenly spaced amount. Such arrays could have a major impact in the field of fiber-optic communications, where it is desirable to send multiple signals, each at a unique wavelength, down a single fiber.

The Bellcore group has also made entire arrays of microlasers that can be turned on and off by light pulses. The arrays could be incorporated in information-processing machines that operate by parallel processing, in which elements of a problem are solved simultaneously rather than in sequence [see "Advanced Computer Architectures," by Geoffrey C. Fox and Paul C. Messina; SCIENTIFIC AMERICAN, October 1987]. Related efforts at Nippon Electric Corporation, the University of California at Santa Barbara and Sandia National Laboratories, among others, are also meeting with success.

Future work in microlasers is likely to take two paths. First, the present generation of microlasers has an output at wavelengths of approximately one micron (infrared radiation). It would be desirable to increase the output wavelength to the region ranging from 1.3 to 1.5 microns, because these wavelengths are transmitted with the fewest losses in an optical fiber and are therefore of potential use in optical communications. To date, microlasers operating in the longer-wavelength region have not worked very well, primarily because of a lack of suitable semiconductor materials that can be grown into a stack of layers forming mirrors for those wavelengths. At present, instead of being five or six microns tall, an efficient microlaser working at 1.5 microns would have to be about 20 microns tall, which would take almost 24 hours to grow precisely using molecular-beam epitaxy. Such a structure would not be very practical, but alternative designs still to be demonstrated may prove sufficient.

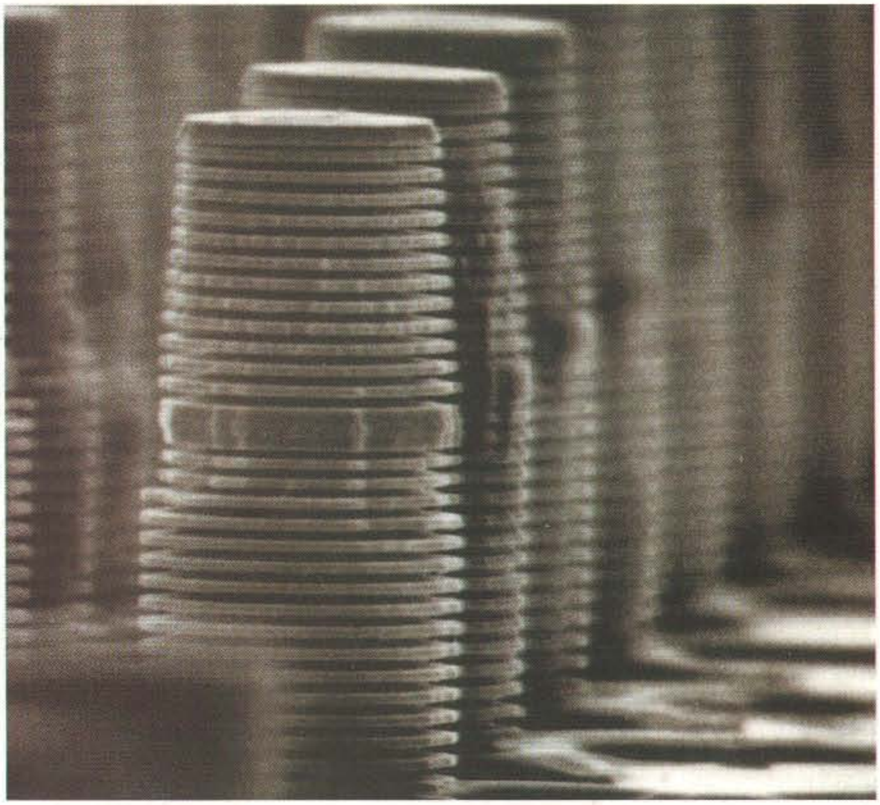
Further miniaturization, which is also just now being explored, is the second likely line of future research. The road will be difficult; we are already close to reaching the small-size limit beyond which the lasers will no longer work. The reason is that the diameter of the laser must be comparable to the wavelength of the radiation it emits; if the diameter is smaller than the wavelength, it does not contain the light effectively. Admittedly, there is an added bonus of working with gallium arsenide: although the laser wavelength is about one micron in air, within the laser itself it is about three times shorter because of the high index of refraction of the semiconductor. At some size smaller than 0.3 micron, however, a microlaser will no longer work.

Why bother to shrink the size of microlasers from a relatively comfortable one micron to, say, a difficult one third of a micron? Further reduction could allow an order of magnitude more information to be communicated and processed with the same power requirements. Second, even more fundamental advantages in power and speed could be attained—advantages that are not predicted by simple scaling.

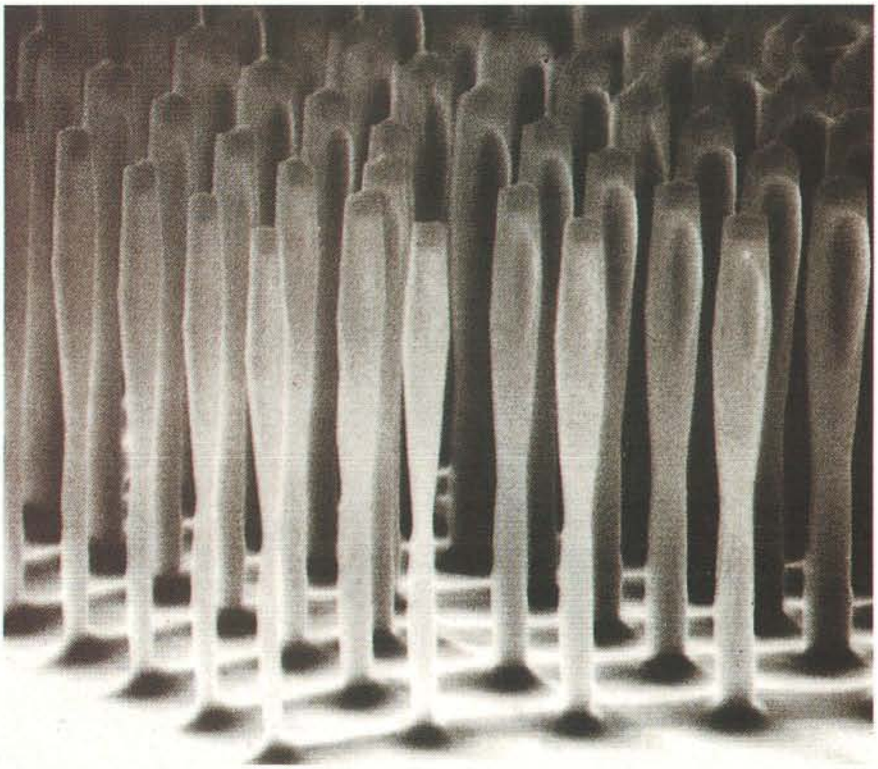
Virtually every laser is inefficient. The first flash of light emitted by the pumped atoms in the amplifying medium goes randomly in all directions. Most of the light does not even strike the laser mirrors, and it is lost. Only a tiny fraction, one part in 10,000 or so, travels in just the right direction to contribute to the working of the laser. It turns out, however, that when the cavity containing the amplifying medium is short and has a small diameter, the walls of the cavity itself, through reflections, can modify the direction in which the light is initially emitted. In a small cavity the light can be emitted only in certain allowed directions and at certain wavelengths. In an extremely small cavity the light is emitted only along the direction of the desired laser beam.

It would therefore appear that microlasers having a diameter of approximately one third of a micron may form the ultimate optoelectronic information-processing device: a small, fast and highly efficient laser requiring low power. Although the prime motivation for building such a device would be for information processing, other applications would undoubtedly be found. Panels of visible-light microlasers could provide efficient lighting, turning the applications full circle from exotic optoelectronic computer technology to a more down-to-earth lamp with which you might read this magazine.

We have experimented to see just how small microlasers can be made and still work. For the purposes of our tests, we pump the atoms in the amplifying



MULTILAYERED STRUCTURE of a microlaser is evident in a scanning electron micrograph. To show the structure, layers of aluminum arsenide have been etched preferentially, with the result that layers of gallium arsenide have larger diameters and project out. Such preferential etching is not done on a working microlaser.



WORLD'S SMALLEST MICROLASERS are about eight microns long and only half a micron in diameter. The authors have successfully tested them with bursts of light.

medium with focused bursts of light instead of electric current. It is easier to fabricate light-pumped devices than electrically pumped ones, and although optically pumped devices are never expected to have any commercial potential, our tests with them have helped convince us (and our managers) that an effort to make electrically pumped cavities of such size is not completely outrageous.

So far we have demonstrated that devices having a diameter as small as one half of a micron can work. Interestingly enough, our tests actually precede any accurate theoretical simulation. Modeling the complex interactions arising from the large number of semiconductor layers would consume huge amounts of computer time and might even require an optical computer!

Perhaps the most exciting aspect of microlaser research is the manner in which it spans the disciplines of fundamental physics, device physics, classical optics, semiconductor optics and optoelectronic and systems integration. Microlasers also challenge investigators' creativity to take advantage of the tremendous applications awaiting discovery. Communications and remote sensing are two areas in which optical systems appear to have fundamental advantages over electronic systems. As machine vision becomes increasingly important, arrays of inexpensive microlasers could allow the technology to be adapted to aid the sight-impaired. We believe the variety of applications of microlasers will be exceedingly wide, but at this point it is only a belief. We look forward to rereading our article 10 years from now to see where we were completely off target and where we were right on.

FURTHER READING

HETEROSTRUCTURE LASERS. H. C. Casey, Jr., and M. B. Panish. Academic Press, 1978.

PHYSICS OF SEMICONDUCTOR DEVICES. Simon M. Sze. John Wiley & Sons, 1981.

THE TECHNOLOGY AND PHYSICS OF MOLECULAR BEAM EPITAXY. Edited by E.H.C. Parker. Plenum Publishing, 1985.

DIMINISHING DIMENSIONS. Elizabeth Corcoran in *Scientific American*, Vol. 263, No. 5, pages 122–131; November 1990.

VERTICAL-CAVITY SURFACE-EMITTING LASERS: DESIGN, GROWTH, FABRICATION, CHARACTERIZATION. J. L. Jewell, J. P. Harbison, A. Scherer, Y. H. Lee and L. T. Florez in *IEEE Journal of Quantum Electronics*, Vol. 27, No. 6, pages 1332–1346; June 1991.

Make, Model and…

A privacy advocate puts license plates on-line

If you drive a car in Massachusetts, Simson L. Garfinkel probably knows who you are. David Lewis of the Massachusetts Registry of Motor Vehicles told a session at the Computers, Freedom and Privacy Conference in San Francisco that the agency is required by law to sell its registration file for the cost of copying. "So how much does it cost?" asked Garfinkel, a computer journalist and technical adept. "What fields does it contain?"

The answer: $77 for a magnetic tape containing nine million registration records with the make, model and year of each car, plus the name and address of the owners, the date of registration and any liens against the vehicle.

Garfinkel hopes to make the file—all two gigabytes or so—available to one and all for searching via computer network as an exercise in freedom of information. A data-base consultant at the conference estimated that a high-end personal computer could process several requests per second from car thieves, stalkers, marketers, the merely curious and other agents of social and economic change. California restricted access to its motor vehicle files several years ago, after an aberrant fan tracked down actress Rebecca Schaeffer through her automobile registration and killed her. But registration and license records are open to the public in most states. So are court records, real-estate title listings and even, in some cases, the files of public gas and electricity companies.

So, what has been protecting our privacy? Mainly time and trouble. In the past, those wishing to search public records either had to pore through stacks of documents or find a mainframe computer. Garfinkel's plan, however, highlights the growing conflict between the presumption of open public records and citizens' desire for privacy. Desktop computers can now assemble a dossier of financial, medical and other information at the touch of a few keys.

Advocates of free access to such information assert that it can be used to lubricate the wheels of commerce, aid in medical care or improve the quality of government. For example, everyday credit-card transactions rely on financial data bases. In some hospitals, physicians can retrieve patients' records in seconds instead of an hour or more (about half the time, paper records arrive too late to be of any use, notes Eunice Little of the American Health Information Management Association). And shortly after traffic citation records became available in Massachusetts, Lewis pointed out, newspaper reports exposed an appeals commission that was letting off up to two thirds of the drunk drivers who appeared before it. Such an investigation would have been virtually impossible without a computerized search.

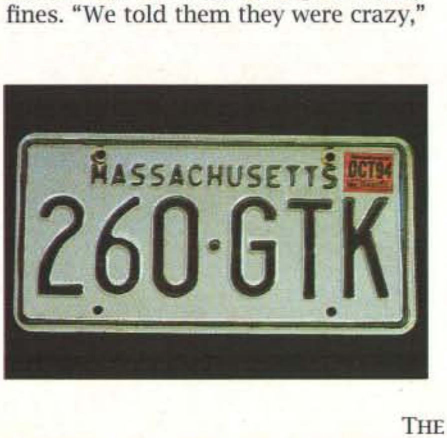
The privacy-minded rebut by pointing out the hazards that accompany easy access to information. Although murders aided by public and private data bases are rare, tales of financial damage are widespread. Indeed, Jack H. Reed, chairman of Information Resource Service Company, a personal-data seller, told the conference audience how he had been denied a mortgage because of a misleading credit report. Insurers, who subscribe to a centralized medical information data base, have been accused of denying coverage to people who have had themselves tested for HIV, even if they test negative, on the theory that being worried enough to take the test implies risky behavior.

These kinds of potential abuses are becoming more important as lawmakers (and private companies) put personal data to uses for which they were never intended. Federal law, for example, supports using motor vehicle information to track parents whose child-support payments are late; a single data-base entry can cause computers to issue a warrant for the alleged deadbeat's car to be seized. In a striking mismatch of crime and punishment, Massachusetts legislators have proposed blocking license renewal for citizens with unpaid library fines. "We told them they were crazy," Lewis notes. If automotive files, containing only name, address, vehicle identification number and a few other bits of information, can spur such controversy, what of medical information? Clinton administration policymakers regard automated medical records as a crucial ingredient in cutting health care costs—Rene C. Kozloff, a project officer at Kunitz and Associates, a health-management information firm, anticipates a "conception to death record" stored on either smart cards or a central data base.

Yet there are minimal controls over the five or six dozen people who may handle those records as a result of a visit to a hospital or clinic. Given the problems that have been caused by disclosure of medical records kept on paper, opening such information to massive, uncontrolled computer searches seems unwise, says Janlori Goldman of the American Civil Liberties Union.

Privacy advocates have been working for nearly 20 years for a so-called Fair Information Practices Act that would give the subjects of public and private data bases power over how personal information on them is used. Although pro-privacy forces have thus far been unsuccessful in the U.S., they have had more luck in Europe. The British enacted "Data Protection" rules in 1984, and a privacy directive for the European Union is in draft form.

British law requires businesses that keep data bases to register them with the Data Protection Registrar, to ask for people's consent before gathering information about them and not to use those data for a purpose different from the one for which they were collected. "Information about others is held in trust" rather than being owned by data-base compilers, says Rosemary Jay, legal adviser for the registrar. Jay has brought court challenges against credit-reporting agencies; she has also had to deal with direct marketers seeking access to the registrar's list of data bases. "Cheeky," she comments. —*Paul Wallich*



LICENSE PLATE in Massachusetts is one of about nine million available on a database. State law requires the motor vehicle agency to sell its license registration file, holding detailed information on every car and driver, to all comers. What a buyer gets (for $77) is a magnetic tape holding data on automobile registrations in Massachusetts: the make, model and year of each car, together with the owner's name and address and other information.

Flat-Panel Displays

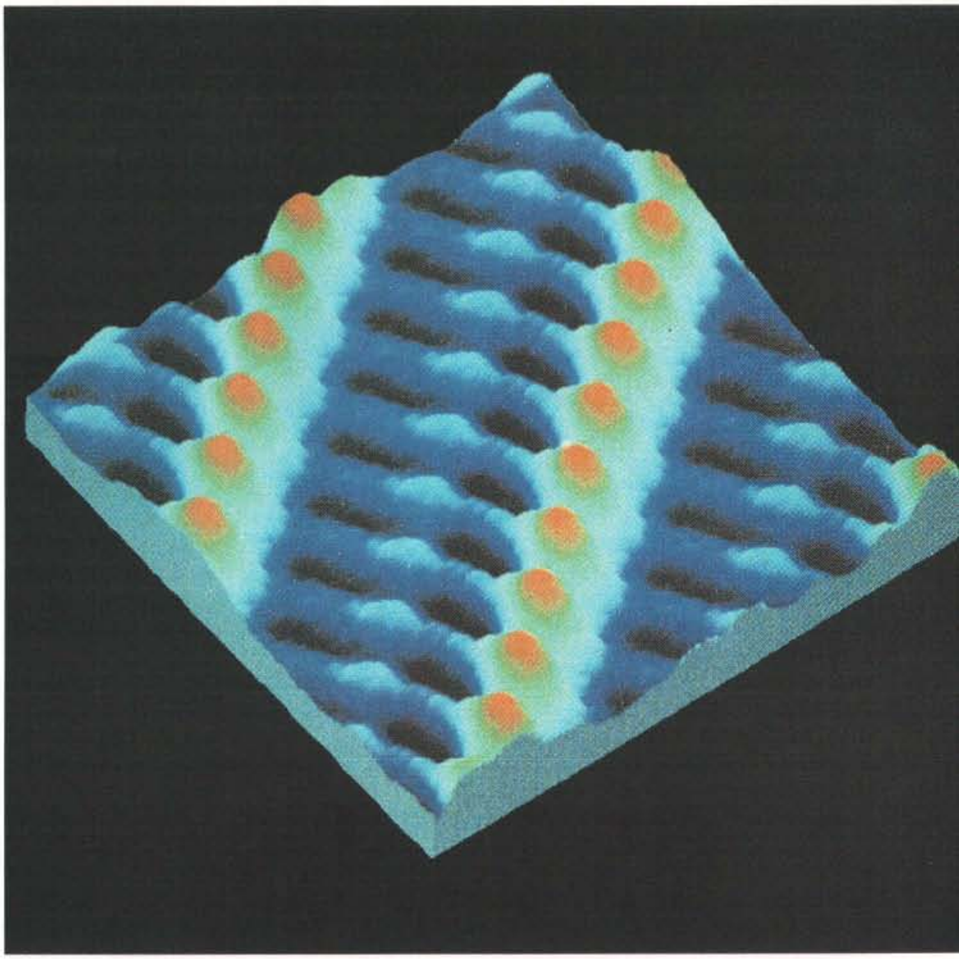
Recent advances in microelectronics and liquid crystals make possible video screens that can be hung on a wall or worn on a wrist

by Steven W. Depp and Webster E. Howard

From the family television set to the computer terminal, the electronic display has become an indispensable way to deliver information. No other medium offers its speed, versatility and interactivity. These attributes are being used to create a wide variety of products that can provide information in any combination of text, graphics, still images or video. Further evolution of this technology will depend, to a great extent, on advances in flat-panel displays.

Although the conventional cathode-ray tube (CRT) remains the dominant display, it has been difficult to modify this technology into a form that is portable, sparing in its use of power and yet capable of producing a superior image. Attempts to squash the CRT into a flat panel have led either to inferior picture quality or to complex designs with excessive manufacturing costs. The latest such attempt replaces the scanning electron gun with an array of tiny electron emitters fabricated on one plate and capped by a second, with the plates kept apart by spacers. But to support the vacuum between the plates, the spacers must be large or numerous; in either case, they tend to obscure the displayed image. Engineers are still seeking a satisfactory and manufacturable solution.

By their nature, flat-panel displays occupy small volume, weigh little and require modest amounts of power. Some can even be written on, like paper. Once a flat panel can be carried about or even worn on one's wrist, like a watch, an individual user will be able to have access to any information, in any place, at any time. Such displays are now at hand. Some airlines use them to provide passengers with armrest movies; in Japan,



LIQUID-CRYSTAL MATERIAL is sealed between two glass plates, one bearing transistors to control the electrodes of each cell, the other bearing color filters and an electrode to complete all circuits. Polarizers in the front and rear complete the array, which is illuminated from behind. Liquid-crystal molecules, fixed to a substrate, appear in the scanning tunneling micrograph above.

STEVEN W. DEPP and WEBSTER E. HOWARD study display technologies—Depp at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y., and Howard at AT&T. Depp is department group manager of entry system technologies. He earned a doctorate in physics from the University of Illinois in 1972 and then worked for five years at the Los Alamos Scientific Laboratory. In 1977 he joined IBM's research laboratory in San Jose, Calif., where he specialized in display technologies. In 1982 he moved to Yorktown Heights. Howard earned a doctorate in physics at Harvard University and joined IBM in 1961. He worked at the Watson Center until 1993, conducting research on semiconductor physics, thin-film electroluminescence and active-matrix liquid-crystal displays.

Reprinted from the March 1993 issue

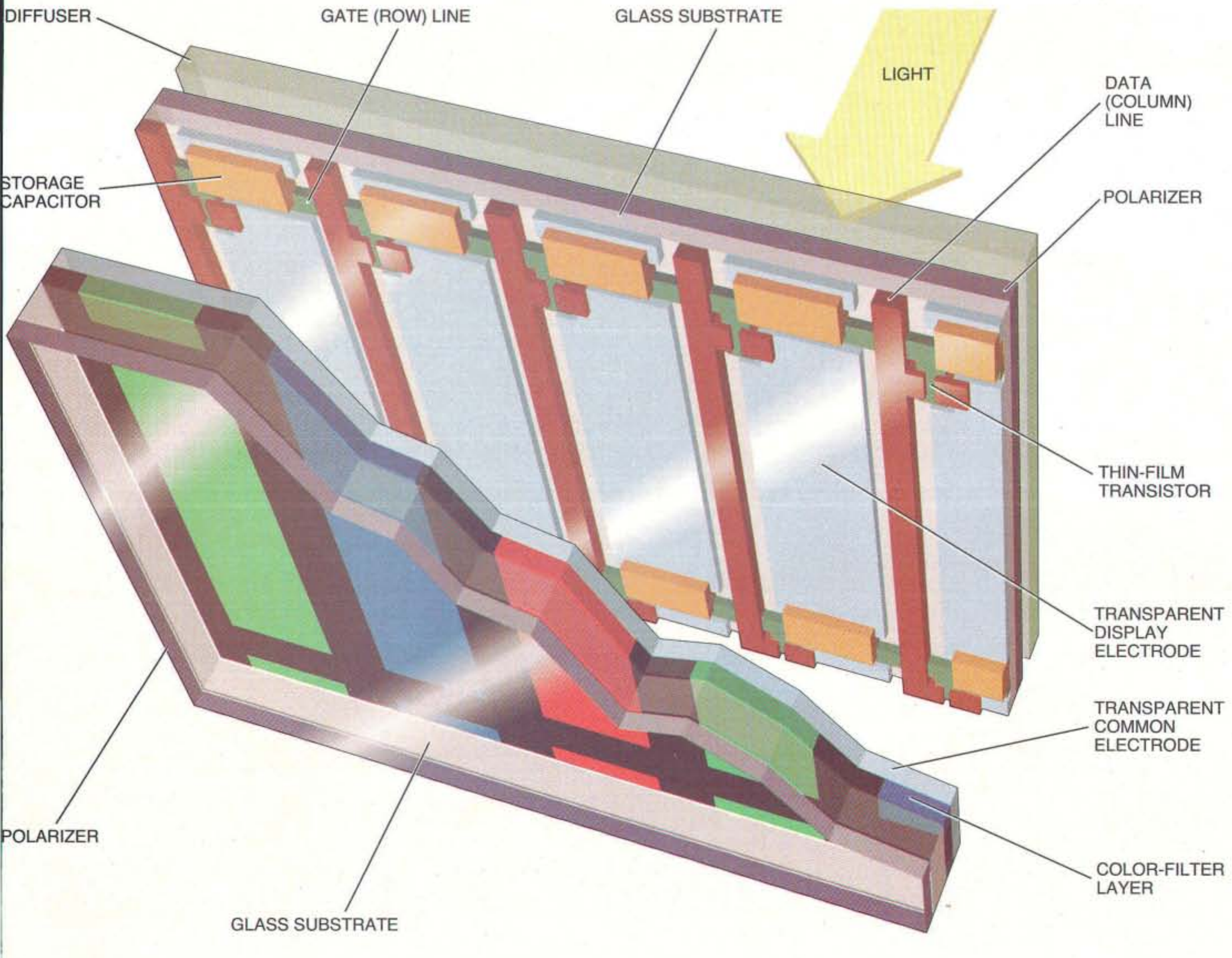
hotels use them in elevators to advertise their restaurants and shops. Within 10 years, high-definition television and dashboard-mounted navigation systems will be commonplace. Flat displays will also spur the development of entirely new products, such as pocket-carried notepads that can store all the memos one might ever write and search them for information keyed to words or dates.

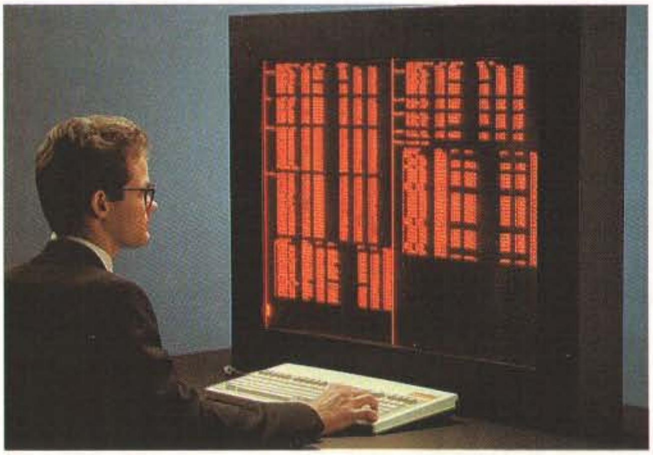
The ideal flat display would portray images with good brightness, sharp contrast, high resolution, quick response, many shades of gray and all the colors of the rainbow. It would also be rugged, long-lasting and inexpensive. The three technologies that have achieved some market acceptance—plasma panels, electroluminescent displays and liquid-crystal displays—all fall short in one or another of these respects, but a fourth just reaching market provides an image that rivals and in some ways surpasses that of the CRT. It is the active-matrix liquid-crystal display (LCD), which currently garners most of the research and development resources and manufacturing investment.

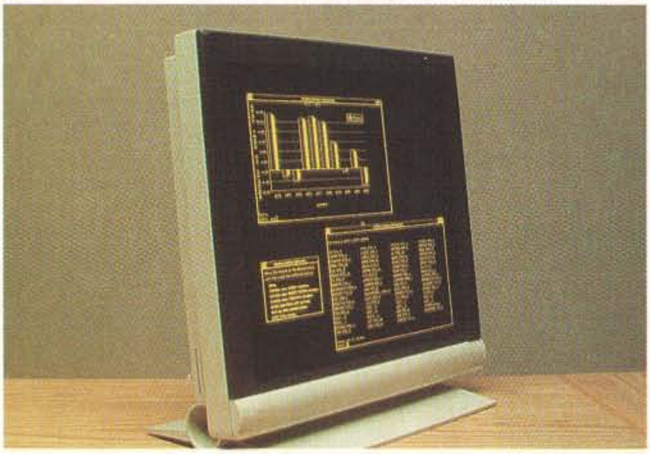
When workers first began to develop flat-panel displays, no one had conceived of active-matrix liquid-crystal devices, and many approaches for making or modulating light in controlled patterns were tried. The first technology to be used in flat panels with high information content employed plasma, or gas discharges. It appeared in the late 1960s.

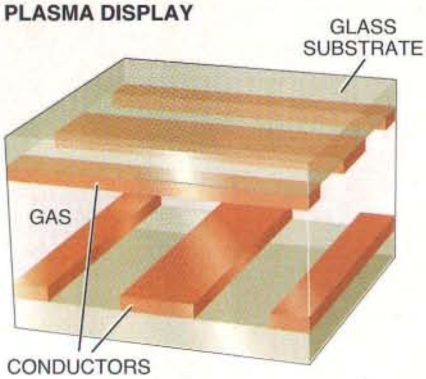
A plasma display begins with two sheets of glass fabricated with a set of parallel ribbons of conductive film. The sheets are placed so that the sets cross. The sheets also enclose a small space filled with a mixture of gases that generally includes neon [*see illustration on next page*]. At any intersection point, a sufficiently large voltage will cause the gas to break down into a plasma of electrons and ions, which glows as it is excited by the current. In effect, one has an array of miniature neon lamps that provide their own light, thus constituting an "emissive" display.

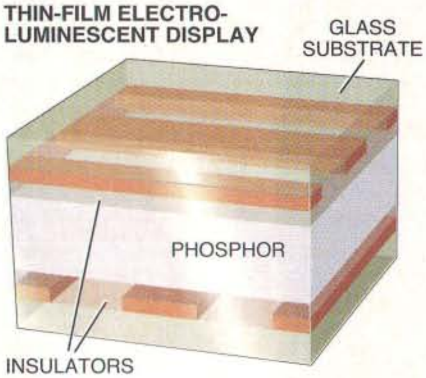
Because the gas ionizes at a well-defined voltage, it is easy to control which intersections light up and which do not. One merely applies a voltage equal to half the firing threshold to a given row and also to its matching column, ensuring that only the crossover point has enough voltage to discharge. By scanning sequentially, a row at a time, and repeating the process at least 60 times per second, one can trick the eye into perceiving a steady image. The image will necessarily be somewhat dim because a given point cannot glow more than a small fraction of the time. Despite this drawback, the design has found some application in portable computers and a handful of other products.











EMISSIVE DISPLAYS glow with their own light by energizing a plasma or a phosphor. Plasma displays (*left*) consist of an array of miniature neon lamps that are discharged by the combination of one voltage from the cell's row and another from its column. Alternating-current designs (*not shown*) use insulated electrodes to retain a charge, creating a memory effect. Thin-film electroluminescence displays (*right*) replace the gas with a solid phosphor film. Emissive displays are rugged but consume too much power for some portable applications and cannot easily provide full color.

An elaboration of this approach achieves a brighter image by employing alternating current. Its inventors, H. Gene Slottow and Donald L. Bitzer of the University of Illinois, found that alternating current would produce an inherent memory effect because a cell that has just fired will briefly retain some charge on its insulated electrodes. When the voltage reverses, this charge will add to any new voltage that is applied and will trigger the discharge again. Because the picture elements (pixels) fire each time the voltage reverses, they emit light for a much larger proportion of the time. Originally, interest in alternating-current plasma centered on its memory, but with today's inexpensive semiconductor memory, the display's main advantages are brightness, sturdiness and reliability. It is particularly prized in military applications.

Most plasma displays are orange. It is not easy to modify the design to produce full color, a feature that consumers have come to expect (more than 80 percent of the computer monitors now sold provide full-color displays). For color, three sources—one emitting red light, the others, blue or green—are required. But it would hardly be practical to achieve these colors by filling each pixel with a different gas. Engineers have therefore emulated the fluorescent tube by getting their light in one form and using phosphorescent coatings, or phosphors, to convert it into several others. One gas, common to the entire array, emits ultraviolet light. This invisible radiation then causes each pixel to glow red, blue or green, depending on its phosphor.

Because plasma displays consume a great deal of power, they appear mainly in applications in which energy efficiency and portability are not required. But such displays can be fashioned into panels measuring as much as 1.5 meters along the diagonal, enough to make a high-definition television screen that might hang on a wall. If color plasma panels can achieve reasonable cost and high reliability, they may be used in such screens.

Another important emissive flat-panel technology is thin-film electroluminescence. Engineers originally hoped to use electroluminescence as a sophisticated source of home lighting. The technology proved to be not nearly efficient enough, however, and the devices ended up in alphanumeric displays, competing with plasma displays.

Like gas-discharge devices, an electroluminescent display resembles a sandwich. A light emitter, or phosphor—typically zinc sulfide doped with manganese—is placed between insulating layers that bear orthogonal electrodes [*see illustration above*]. When the voltage exceeds a well-defined threshold, the emitter breaks down and conducts current. The current then excites the manganese ions, which give off a yellow light, anal-ogous to the glowing neon of the plasma display.

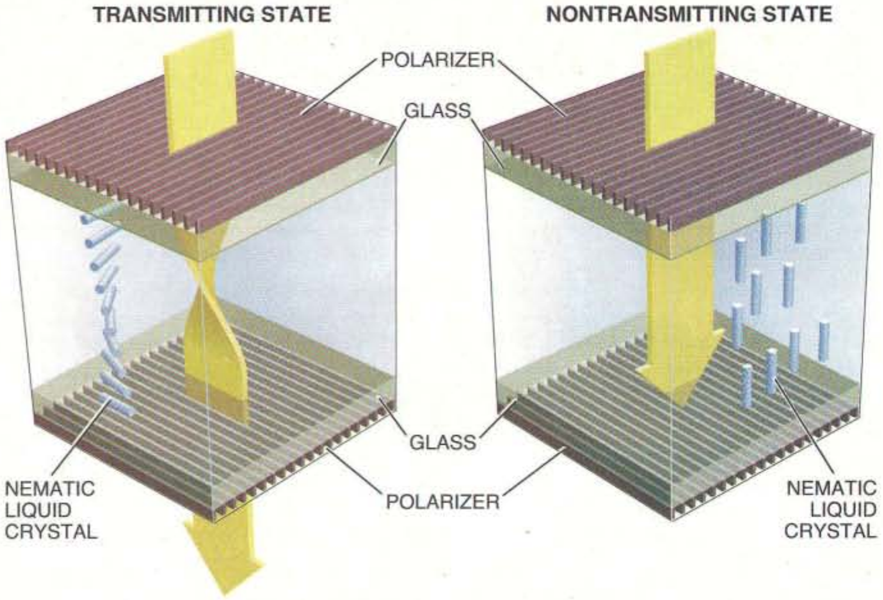
Although such displays are also quite durable, they suffer from two drawbacks: they cannot provide full color, and they consume almost as much energy as do plasma displays. Full color requires a good blue-emitting phosphor, but none having suitable brightness, efficiency and longevity has yet been found. Full color also requires a smoothly varying gray scale, and this has been difficult to control. In addition, the efficiency of such panels declines as the number of picture elements increases. Each element functions as a capacitor, and significant amounts of energy are needed simply to charge and discharge them all. This problem militates against using high-information-content electroluminescent panels in battery-powered applications.

The primary nonemissive technology makes use of the electro-optic properties of a class of organic molecules known as liquid crystals (LCs). As the name implies, liquid crystals are compounds that flow like liquids but have a crystalline order in the arrangement of their molecules. This phase of matter seems improbable, but it is not so rare as one might think. It has been estimated that an organic chemist randomly synthesizing compounds would produce molecules with liquid crystalline properties about once in every 1,000 experiments. These molecules have been known for about 100 years. They appear in many forms, from cell membranes to soap scum.

The most important of the LCs are the nematic, or threadlike, compounds. Their rod-shaped molecules are free to move with respect to one another, yet subtle intermolecular forces tend to keep their long axes aligned. The effect somewhat resembles the schooling of fish. One can set the direction of the molecules' alignment by placing them in an electric field or close to a specially prepared surface. By controlling the alignment, one effectively controls the optical properties as well—above all, the ability of the material to affect the transmission of light.

In one popular configuration, a liquid-crystal compound is sandwiched between glass sheets having different patterns of molecular alignment [*see illustration on this page*]. To construct such a "twisted nematic" LC, fabricators coat the inner face of each sheet with a transparent, electrically conductive film such as indium tin oxide. They then add a thin layer of an organic polymer and finally brush the polymer in a desired direction. Brushing aligns the polymer by combing its chains or making tiny scratches, or both. The surface effects communicate themselves to the adjacent liquid-crystal molecules through a mechanism that remains obscure. The twist is achieved by placing the glass sheets so that their preferred orientations run at right angles to one another.

If polarized light is directed through the liquid-crystal cell, the direction of polarization will tend to follow the twist. It therefore emerges from the sandwich rotated by 90 degrees. Crossed polarizers on either side of the cell will then pass light rather than blocking it, as would be expected if the cell were not present. This is the "on," or transmitting, state. The "off," or nontransmitting, state occurs when an electric field is applied via the two transparent conductors. The field arranges the liquid-crystal molecules with their long axes parallel to the field lines. This realignment removes the twisted configuration, and thus the cell no longer rotates the incoming polarization. The cell then has no significant optical effect, and the crossed polarizers block the light as usual. In an actual cell, the transmission varies smoothly with voltage.



LIGHT-TWISTING SHUTTER is used in most nonemissive flat-panel displays. Threadlike, or nematic, liquid-crystal molecules are arranged so as to rotate the polarization of light. Nonenergized pixels enable the light to pass a second polarizer (*left*). When a voltage is applied, the twist is disrupted, and the light is blocked (*right*).

Thus far we have described a single pixel. A complete liquid-crystal display incorporates many of them by putting the row electrodes on one glass substrate and the column electrodes on the other; each crossover point defines the address of a pixel. Because the light comes from a reflector or light source, the matrix operates as an array of tiny, electronically controlled shutters. The simplicity of design and the modest voltage and power requirements of this "passive matrix" LCD have made it the best-selling flat-panel technology today.

Yet the simplicity of this approach places an inflexible limit on its performance. As Paul M. Alt and Peter Pleshko of IBM have demonstrated, one can improve resolution only by sacrificing contrast. The fault lies in cross talk, an inherent feature of the system by which pixels are energized. Here is how it happens. A voltage is applied to a single row, and the column voltages are adjusted to produce a large combined voltage across the selected pixels in that row. Unselected pixels receive a smaller total voltage [*see illustration on next page*].

The addressing continues as the next row down is energized, a process that sweeps from top to bottom in about one sixtieth of a second, then repeats for the next frame. During the depiction of a frame of video, each selected pixel receives a large voltage pulse when its row is activated; unselected pixels receive moderate voltages, and all pixels receive a series of smaller cross-talk pulses when the remaining rows are addressed.

Because cross talk rises with the number of rows, the difference in effective voltage between selected and unselected pixels declines as the size of the array increases. For arrays having 240 rows—half as many as a typical television screen—the difference comes to only 6.7 percent. For a nematic cell twisted through 90 degrees, an acceptable contrast would require a difference in voltage of at least 50 percent.

There are three strategies for getting around the trade-off between contrast and resolution. First, one can try to make the transmission-voltage curve steeper, so that small differences in voltage yield large changes in transmission. "Supertwisted nematic" LCDs do this by twisting the liquid crystal through 180 degrees or more. Second, one can use liquid crystals that exhibit a memory effect, enabling one to address many rows without sacrificing contrast. Ferroelectric liquid crystals have this property, and they have been assembled into displays of 1,000 rows and more. But they are somewhat slow. Moreover, because they have only two stable transmission states, they cannot easily reproduce various shades of gray. These drawbacks render the passive-matrix approach less suitable for the display of realistic images and video.

The third and most radical approach separates the functions of addressing and transmitting, so that each may be optimized. Such "active matrix" addressing employs an array of transistors, each of which activates a single pixel. Each pixel receives a voltage from its column line only when its own transistor is switched on. The other rows can be addressed when the transistor switches off; meanwhile the pixel holds the voltage it had been given. Because this method effectively isolates the pixel from cross talk, the number of addressable rows can be extremely large. In addition, the technique can easily provide color. Pixels and their respective transistors are simply grouped into triads in which each pixel has a filter for one of the three primary colors [*see illustration on page 71*].

The entire display is rewritten after all the rows have been addressed. Such refreshing, as it is called, prevents the distortions that might otherwise appear as charge gradually leaks from the liquid-crystal cells, reducing their voltages and changing their transmission of light. More important, by refreshing every sixtieth of a second, the pixels can track video and other rapid screen changes.

Active-matrix displays closely resemble dynamic random-access memory (DRAM) chips. Both are complex integrated circuits that store charge in a million or so discrete locations, each one under the control of a single transistor. Yet a computer reads a DRAM chip a row at a time by sensing the charge at each location, whereas the human eye in effect reads the entire display in parallel.

Another difference is that DRAM stores digital data; active-matrix displays store analog data. In the former, each cell must be either mostly "on" or "off"; in the latter, cell voltages must vary along a gray-scale continuum. This feature requires the display to "encode" far more information than a DRAM chip of similar size can manage.

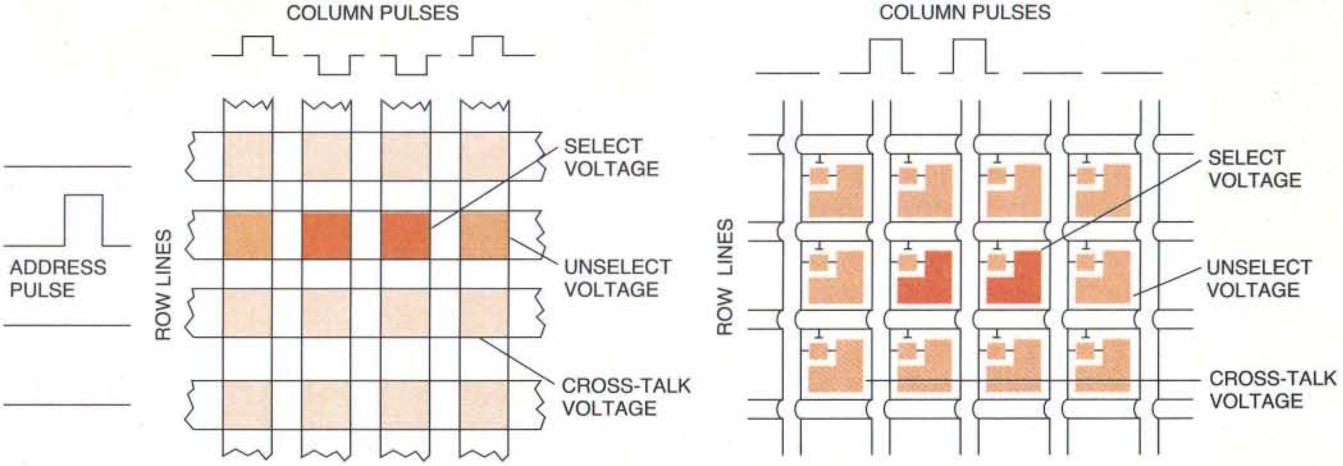
It has been possible to meet this constraint by careful design and controlled manufacturing tolerances. Fortunately, these tolerances need apply only on a small scale, from pixel to pixel. The reason is that although the eye is sensitive to the local variations produced at the edges of objects, it hardly notices the variations that build gradually as it scans from one side of the screen to the other.

An active-matrix display is controlled by thin-film transistors (TFTs), a technology that has come into its own only in the past decade. Such transistors derive from solar-cell manufacturing technology and can be made in large arrays at comparatively low cost. The only drawback—a lower current-carrying capacity—hardly matters in display applications because very little current is needed to control a liquid crystal.

A thin-film transistor uses deposited layers of different materials to constitute the semiconductor, the insulators and the electrodes. As in conventional transistors, two terminals conduct current, and the third switches the transistor on and off. But conventional transistors form most of these parts on the surface of a single, semiconducting crystal, whose electrical properties are modified in particular regions by the addition of charge-donating atoms, called dopants. The TFT can be fabricated on virtually any surface, including inexpensive glass.

It took about 20 years to turn the idea of the TFT into a commercial product. When Paul K. Weimer of RCA invented the technology in 1962, experts expected that it would find application in conventional electronic circuits. But other electronic devices quickly superseded it, discouraging most researchers from working on TFT technology. Only in 1974 did T. Peter Brody and his colleagues at Westinghouse demonstrate that TFTs might be used as switches for liquid-crystal displays. Even that application proved difficult at the time because the available materials and processes could not produce large arrays of stable, defect-free devices.

Researchers advanced by trying a variety of semiconductors. Polycrystalline silicon became the first to find its way into a commercial product in 1984, when Seiko-Epson used it in a pocket color television set with a remarkably attractive two-inch screen. But the display had to be fabricated with expensive integrated-circuit processes and high-temperature materials. Already it was apparent that the future lay with another material, amorphous silicon. In 1979 P. G. LeComber and his colleagues



PASSIVE VERSUS ACTIVE: passive liquid-crystal matrices (*left*) give all pixels in a given row an address pulse, then add or subtract a column voltage. Unselected cells thus accumulate an average "cross-talk voltage" that rises with the number of rows, producing a trade-off between contrast and resolution. Active matrices (*right*) avoid this problem by using an array of thin-film transistors to address the liquid crystals. While the transistors are switched on, the associated liquid-crystal cells are charged to the voltages on the column lines, also known as data lines. When the transistors are turned off, the voltages are held constant. In this way, an image can be written row by row and maintained until the next "refresh cycle."

at the University of Dundee had the insight to see that amorphous silicon, already being developed for inexpensive solar cells, could be suitable for TFTs in displays.

There are several key processes for fabricating the TFT array. First, the substrate is formed of glass that has been purified of alkali metals, which might otherwise contaminate the transistors or liquid crystal. The molten glass is poured out, in a finely controlled process developed by Corning, to produce a sheet so flat that, were it scaled to the size of a football field, its thickness would vary by less than the breadth of a pencil.

Later, the semiconductor layer is deposited by a plasma process using silane gas, SiH_4, under low pressure. An electrical discharge ionizes the gas and breaks the molecules into fragments, which condense on the glass as a random network of silicon, rich in hydrogen. The hydrogen is crucial because it ties up broken bonds that might otherwise trap electrons and disrupt the semiconductor. Finally, metal electrodes, insulators and other elements in the TFT are deposited in a manner similar to that used in fabricating integrated circuits. Here, however, a much larger area must be covered.

About two dozen companies, mostly in Japan, developed TFT/LCD technology during the mid-1980s. By the end of the decade, IBM/Toshiba, Sharp, Hitachi and others were demonstrating excellent image quality in color displays spanning diagonals of 10 inches or more. For the first time, a flat-panel display could match a cathode-ray tube in contrast, brightness and color saturation (or vividness). Indeed, when the two kinds of display are placed side by side, most observers prefer the crisp, undistorted, jitter-free image of the TFT/LCD.

Although the active-matrix display is meeting the standards of image quality set by the CRT, it costs several times as much to produce. Further refinements in manufacturing technology are required; to develop the necessary experience, manufacturers will concentrate, in the beginning, on the many emerging applications for which a flat display is indispensable.

Among such applications are tiny televisions and videotape recorders, portable—even wearable—computers, dashboard-mounted screens for automobile navigation systems, avionics displays for military, commercial and space applications, as well as electronic books, notebooks and clipboards. Soon the list will include still more exotic products, such as shopping-cart displays and products for the emerging industry of virtual-reality entertainment.

WALL-HUNG TELEVISION in the 1990 movie *Total Recall* depicts life in the year 2084, yet comparable products should reach the market within a decade. When not presenting live video, such a panel might revert to the pattern of the surrounding wallpaper, display works of art from a digital library or serve as an interface with a computer network.

By 1995 such applications are expected to form a market of between $4 billion and $5 billion. That scale of production should eventually reduce unit costs enough to make flat-panel displays become competitive with CRTs in television, desktop computing and other traditional markets. At the end of this decade there should be more revenue from flat panels than from CRTs.

We predict that as flat panels continue to evolve toward higher resolution and reduced power and weight, they will increasingly replace paper. In addition to their flexibility, they offer the advantage of responding to touch or stylus. These qualities all tend to simplify the most cumbersome element in modern electronics, the interface, enabling the user to take full advantage of the computer itself.

One such display, with data storage, might replace a filing cabinet full of forms. Entirely new uses will also arise. When one is not using one's screen as a television or as a computer monitor, it may disappear, like a chameleon, into the wallpaper pattern or serve as an art gallery or photograph album. We can easily foresee a time when even this magazine will be an electronic publication viewed on a personal flat-panel display, enlivened by video graphics and backup information available at a touch of the screen.

FURTHER READING

FLAT-PANEL DISPLAYS AND CRTS. Edited by Lawrence E. Tannas, Jr. Van Nostrand Reinhold Company, 1985.

LIQUID CRYSTAL TV DISPLAYS: PRINCIPLES AND APPLICATIONS OF LIQUID CRYSTAL DISPLAYS. E. Kaneko. D. Reidel Publishing Company, 1987.

LIQUID CRYSTALS: APPLICATIONS AND USES, Vol. 1. Edited by Birendra Bahadur. World Scientific, Singapore, 1990.

FLAT PANEL DISPLAYS USING AMORPHOUS AND MONOCRYSTALLINE SEMICONDUCTOR DEVICES. K. Suzuki in *Amorphous and Microcrystalline Semiconductor Devices: Optoelectronic Devices.* Edited by Jerzy Kanicki. Artech House, Boston, 1991.

THIN-FILM-TRANSISTOR/LIQUID CRYSTAL DISPLAY TECHNOLOGY: AN INTRODUCTION. W. E. Howard in *IBM Journal of Research and Development,* Vol. 36, No. 1, pages 3–10; January 1992.

When Cells Divide

Making space for the next wave of wireless communications

If Benjamin were in Westwood lounging around his parents' pool today, the word whispered in his ear would be "wireless." The wave that has rapidly but calmly lifted radio telecommunications from a niche technology to an $8-billion, 11-million-customer industry in just 10 years is starting to break. It has met another wave: the swell of digital networks as computers evolve from tools of calculation to portals of communication. Caught in the spray, cellular companies are thrashing about to get atop this confluence.

For the industry to thrive, it must simultaneously become more competitive and more cooperative. Simply lowering the price of admission to wireless networks is not enough; the diverse equipment and services must all work together as well. With the right balance, the wireless market could quintuple in the U.S. over the next decade, predict market analysts at Arthur D. Little. That event would bring personal, portable telephone services to 60 million people and their computers by 2005.

Substantial technological hurdles must be cleared if that is to happen. The first is capacity: the analog radio infrastructure in place simply does not have enough channels to support an explosion of demand. The FCC's decision in September 1993 to auction off 160 megahertz of bandwidth for personal communications services—more than three times the portion of the spectrum currently allotted to cellular telephone systems—is but a temporary fix. For the FCC also decided to allow existing mobile telephone companies to purchase only small plots of that valuable real estate; the rest will go to new entrants, up to seven per city.

As the resulting competition drives prices down, rising consumer expectations seem likely to force capital investment up. In order to entice customers not reared on the fuzzy, unreliable connections typical of the analog Advanced Mobile Phone Service (AMPS) standard, cellular companies must be able to offer not only land-line-quality speech transmission but also reliable connections for a new breed of fast computer modems.

Indeed, as Benjamin L. Scott, chief operating officer for Bell Atlantic Mobile, notes, "the growth potential for wireless data service is enormous. In our markets alone, $1.3 billion is up for grabs." Cellular firms realized long ago that to meet all these demands, they would have to replace AMPS with a digital standard. The prospect of being trapped with shrinking margins and no way to add customers has simply added a note of urgency to the debate over what that standard should be.

In its rush to develop digital services, however, the cellular industry has made several false starts. GTE Mobilnet and Ameritech, among others, have announced with much fanfare a service they have begun testing: Cellular Digital Packet Data. By stuffing chunks of your data into the pauses in other people's conversations, the service should provide wireless connection to the Internet at speeds up to 19,200 bits per second. But if all-digital wireless service is introduced over the next two years as expected, such analog-based services will succumb to obsolescence.

The cellular industry's first attempt at a digital standard may suffer the same fate. The Telecommunications Industry Association adopted a scheme called Time Division Multiple Access (TDMA) in 1989. TDMA increases the capacity of analog cellular systems up to six times by chopping conversations into short segments and interleaving pieces from several conversations into each digital channel.

Ten large cellular telephone companies signed on as early adopters of TDMA. But within a year the standard was challenged by a rival method proposed by Qualcomm, a San Diego startup. The company had patented a technique called Code Division Multiple Access (CDMA), which it claimed could increase by 10 to 20 times the capacity provided by AMPS, while delivering better quality than TDMA.

Qualcomm tries to do this by avoiding the use of channels. Instead CDMA dumps all the transmissions sent within a cell into one wide band. It keeps individual signals separate by assigning each one a computer-generated code. That code is then used to manipulate data or digitized speech mathematically so that its bits are spread evenly throughout the spectrum, where they mingle with bits from up to 61 other conversations. A base station, portable telephone or laptop modem receiving a CDMA call can use the same code to unscramble an incoming message. All other transmissions, scrambled with different codes, look like ordinary static to the receiver and are filtered out.

By reducing interference to incoherent noise (as opposed to a coherent conversation), CDMA can use the same wide band of spectrum in every cell. TDMA and AMPS, in contrast, must carefully assign channels so that no frequency is ever used in two adjacent cells. Aside from wasting spectrum, this forces a caller's telephone to switch channels abruptly when the caller crosses a cell boundary. Such hand-offs are the main source of dropped calls. With no channels to switch, two CDMA base stations can talk to the same telephone at once, allowing "soft" hand-offs. And because CDMA systems can fill the spectrum with more information, they can accommodate more callers.

Qualcomm has tacked other innovations onto its standard to boost capacity further. Whereas TDMA always digitizes speech by sampling it 8,000 times per second, CDMA uses a variable-rate digitizer that sends just 1,000 bits per second during the 60 percent of a typical conversation that a person spends listening or thinking. The extra airwaves can be used by other calls. A so-called rake receiver turns the bane of radio communications, multipath distortion, to its advantage. The problem occurs when a signal, having bounced off buildings or hills, arrives from several directions at slightly different times. In television, this problem causes ghost images. The rake receiver watches for such reflections, picks the three strongest and combines them to produce a clearer signal. Finally, built-in power control lets base stations instruct portable telephones to turn their transmission power up or down to avoid fading out of range or overwhelming other signals.

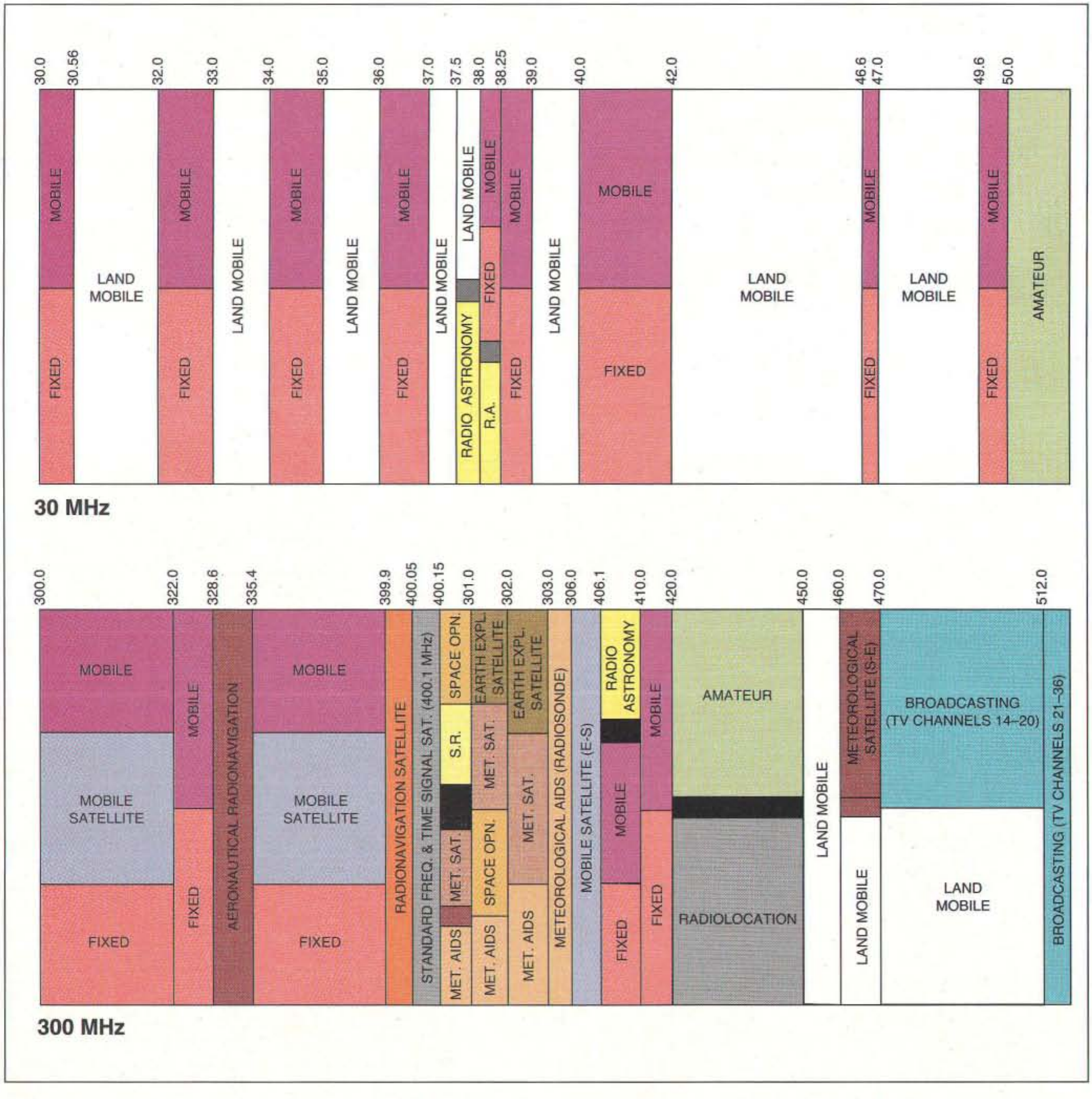
It took three years of testing, but Qualcomm finally convinced the industry to canonize CDMA as a second digital standard in July 1993. Since then, the technique has steadily gained support, garnering commitments from three of the biggest cellular service providers, with another three expected to sign up soon. U.S. West has begun installing

CDMA equipment in the Seattle area, according to a company spokesperson, and planned to offer the service to customers late in 1994. Yet the majority of carriers are not expected to go digital until well into 1995.

By that time, observes Gregory Pottie, a wireless technology researcher at the University of California at Los Angeles, the state of the art may have advanced considerably. "I think there is another factor of four or five to be gained in capacity" above what CDMA promises, he says. Clever use of multiple antennas in each telephone can reduce distortion, for example. Engineers at U.C.L.A. and the Georgia Institute of Technology are testing improved coding schemes that adapt to changing interference conditions to help reduce errors. And AT&T is working on advanced speech-compression algorithms to halve the amount of data needed for land-line quality.

Longer-term gains may come from research into ways of canceling out interference altogether. "The base station is receiving signals from every user anyway," Pottie points out. "Why shouldn't it process this information to wipe out unwanted interference between callers?" This need not require a supercomputer, he asserts. "Typically, there are a few users who dominate the interference; you cancel just the worst offenders."

Combined thoughtfully, these innovations could open enough airspace for 150 million wireless customers in the U.S., Pottie estimates. If the cellular industry can draw half that many away from traditional copper-wire telephone companies and future optical-fiber and coaxial cable services, it will have done well for itself indeed. —*W. Wayt Gibbs*



FREQUENCY ALLOCATIONS made by the Federal Communications Commission in 1987 are shown in part. The FCC decided in 1993 to triple the number of frequencies that can be used for cellular telephone (land mobile) services.

The Computer for the 21st Century

Specialized elements of hardware and software, connected by wires, radio waves and infrared, will be so ubiquitous that no one will notice their presence

by Mark Weiser

The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.

Consider writing, perhaps the first information technology. The ability to represent spoken language symbolically for long-term storage freed information from the limits of individual memory. Today this technology is ubiquitous in industrialized countries. Not only do books, magazines and newspapers convey written information, but so do street signs, billboards, shop signs and even graffiti. Candy wrappers are covered in writing. The constant background presence of these products of "literacy technology" does not require active attention, but the information to be transmitted is ready for use at a glance. It is difficult to imagine modern life otherwise.

Silicon-based information technology, in contrast, is far from having become part of the environment. More than 50 million personal computers have been sold, and the computer nonetheless remains largely in a world of its own. It is approachable only through complex jargon that has nothing to do with the tasks for which people use computers. The state of the art is perhaps analogous to the period when scribes had to know as much about making ink or baking clay as they did about writing.

The arcane aura that surrounds personal computers is not just a "user interface" problem. My colleagues and I at the Xerox Palo Alto Research Center think that the idea of a "personal" computer itself is misplaced and that the vision of laptop machines, dynabooks and "knowledge navigators" is only a transitional step toward achieving the real potential of information technology. Such machines cannot truly make computing an integral, invisible part of people's lives. We are therefore trying to conceive a new way of thinking about computers, one that takes into account the human world and allows the computers themselves to vanish into the background.

Such a disappearance is a fundamental consequence not of technology but of human psychology. Whenever people learn something sufficiently well, they cease to be aware of it. When you look at a street sign, for example, you absorb its information without consciously performing the act of reading. Computer scientist, economist and Nobelist Herbert A. Simon calls this phenomenon "compiling"; philosopher Michael Polanyi calls it the "tacit dimension"; psychologist J. J. Gibson calls it "visual invariants"; philosophers Hans Georg Gadamer and Martin Heidegger call it the "horizon" and the "ready-to-hand"; John Seely Brown of PARC calls it the "periphery." All say, in essence, that only when things disappear in this way are we freed to use them without thinking and so to focus beyond them on new goals.

The idea of integrating computers seamlessly into the world at large runs counter to a number of present-day trends. "Ubiquitous computing" in this context does not mean just computers that can be carried to the beach, jungle or airport. Even the most powerful notebook computer, with access to a worldwide information network, still focuses attention on a single box. By analogy with writing, carrying a superlaptop is like owning just one very important book. Customizing this book, even writing millions of other books, does not begin to capture the real power of literacy.

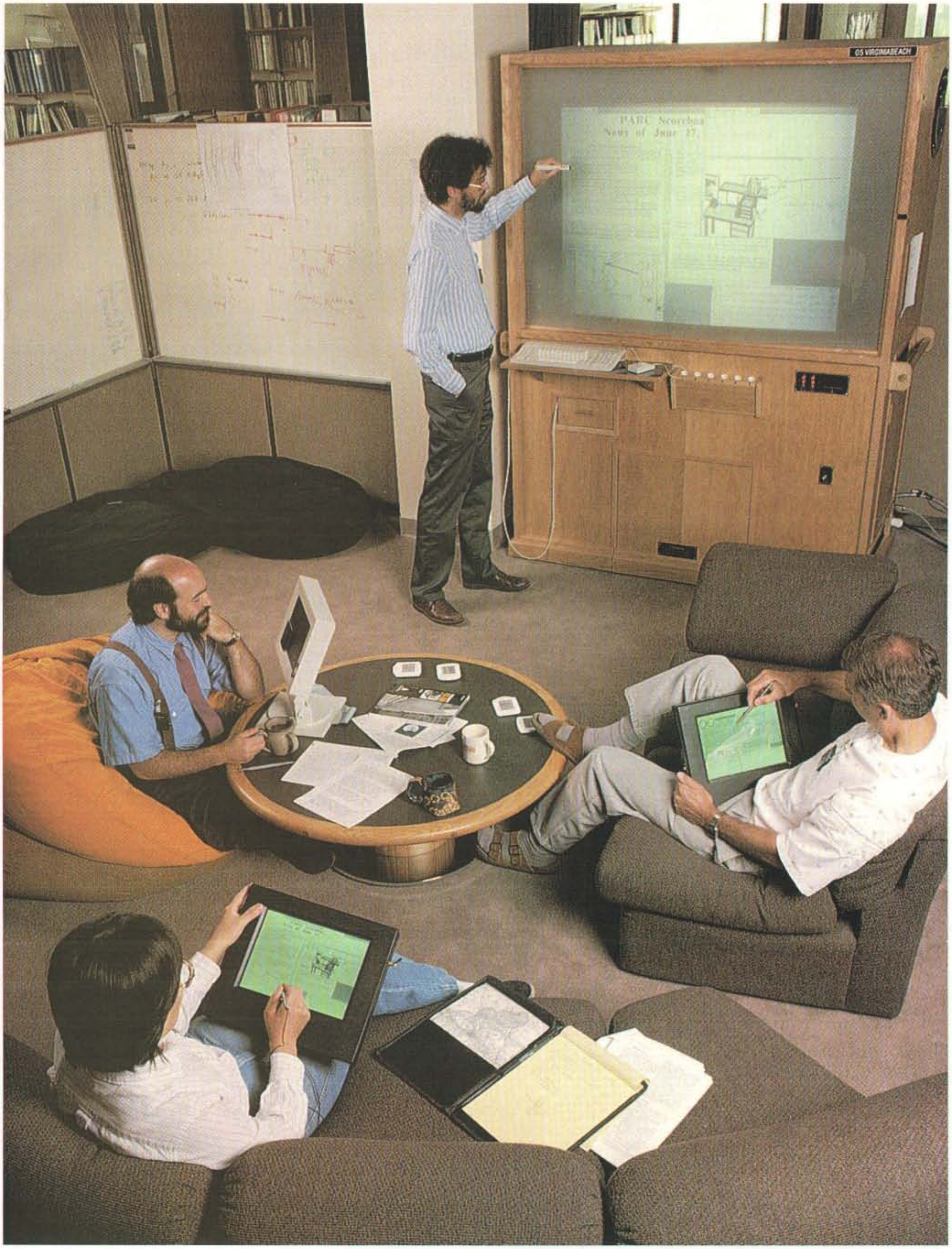
Furthermore, although ubiquitous computers may use sound and video in addition to text and graphics, that does not make them "multimedia computers." Today's multimedia machine makes the computer screen into a demanding focus of attention rather than allowing it to fade into the background.

Perhaps most diametrically opposed to our vision is the notion of virtual reality, which attempts to make a world inside the computer. Users don special goggles that project an artificial scene onto their eyes; they wear gloves or even bodysuits that sense their motions and gestures so that they can move about and manipulate virtual objects. Although it may have its purpose in allowing people to explore realms otherwise inaccessible—the insides of cells, the surfaces of distant planets, the information web of data bases—virtual reality is only a map, not a territory. It excludes desks, offices, other people not wearing goggles and bodysuits, weather, trees, walks, chance encounters and, in general, the infinite richness of the universe. Virtual reality focuses an enormous apparatus on simulating the world rather than on invisibly enhancing the world that already exists.

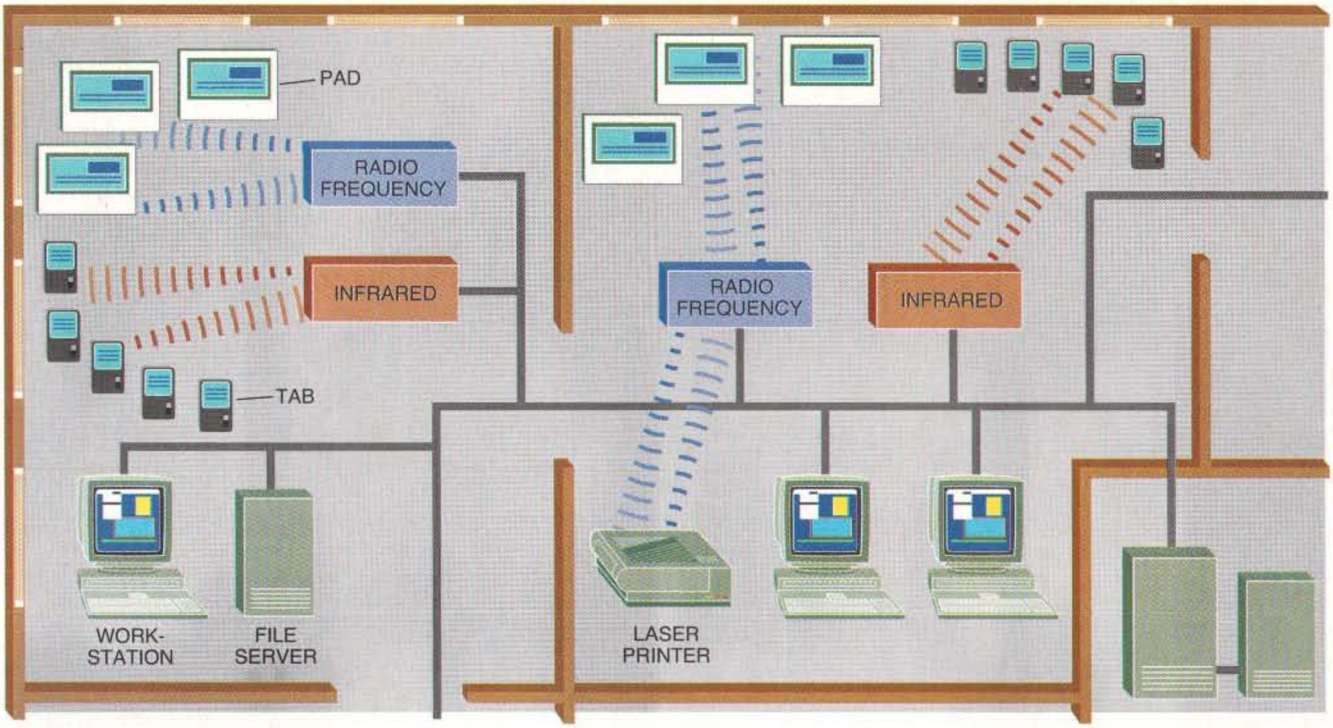
Indeed, the opposition between the

MARK WEISER is head of the Computer Science Laboratory at the Xerox Palo Alto Research Center. He is working on the next revolution of computing after workstations, variously known as ubiquitous computing or embodied virtuality. Before working at PARC, he was a professor of computer science at the University of Maryland; he received his Ph.D. from the University of Michigan in 1979. Weiser also helped found an electronic publishing company and a video arts company and claims to enjoy computer programming "for the fun of it." His most recent technical work involved the implementation of new theories of automatic computer memory reclamation, known in the field as garbage collection.

Reprinted from the September 1991 issue



UBIQUITOUS COMPUTING begins to emerge in the form of live boards that replace chalkboards as well as in other devices at the Xerox Palo Alto Research Center. Computer scientists gather around a live board for discussion. Building boards and integrating them with other tools has helped researchers understand better the eventual shape of ubiquitous computing. In conjunction with active badges, live boards can customize the information they display.



WIRED AND WIRELESS NETWORKS link computers and allow their users to share programs and data. The computers pictured here include conventional terminals and file servers, pocket-size machines known as tabs and page-size ones known as pads. Future networks must be capable of supporting hundreds of devices in a single room and must also cope with devices—ranging from tabs to laser printers or large-screen displays—that move from one place to another.

notion of virtual reality and ubiquitous, invisible computing is so strong that some of us use the term "embodied virtuality" to refer to the process of drawing computers out of their electronic shells. The "virtuality" of computer-readable data—all the different ways in which they can be altered, processed and analyzed—is brought into the physical world.

How do technologies disappear into the background? The vanishing of electric motors may serve as an instructive example. At the turn of the century, a typical workshop or factory contained a single engine that drove dozens or hundreds of different machines through a system of shafts and pulleys. Cheap, small, efficient electric motors made it possible first to give each tool its own source of motive force, then to put many motors into a single machine.

A glance through the shop manual of a typical automobile, for example, reveals 22 motors and 25 solenoids. They start the engine, clean the windshield, lock and unlock the doors, and so on. By paying careful attention, the driver might be able to discern whenever he or she activated a motor, but there would be no point to it.

Most computers that participate in embodied virtuality will be invisible in fact as well as in metaphor. Already computers in light switches, thermostats, stereos and ovens help to activate the world. These machines and more will be interconnected in a ubiquitous network. As computer scientists, however, my colleagues and I have focused on devices that transmit and display information more directly. We have found two issues of crucial importance: location and scale. Little is more basic to human perception than physical juxtaposition, and so ubiquitous computers must know where they are. (Today's computers, in contrast, have no idea of their location and surroundings.) If a computer knows merely what room it is in, it can adapt its behavior in significant ways without requiring even a hint of artificial intelligence.

Ubiquitous computers will also come in different sizes, each suited to a particular task. My colleagues and I have built what we call tabs, pads and boards: inch-scale machines that approximate active Post-it notes, foot-scale ones that behave something like a sheet of paper (or a book or a magazine) and yard-scale displays that are the equivalent of a blackboard or bulletin board.

How many tabs, pads and board-size writing and display surfaces are there in a typical room? Look around you: at the inch scale, include wall notes, titles on book spines, labels on controls, thermostats and clocks, as well as small pieces of paper. Depending on the room, you may see more than 100 tabs, 10 or 20 pads and one or two boards. This leads to our goal for initially deploying the hardware of embodied virtuality: hundreds of computers per room.

Hundreds of computers in a room could seem intimidating at first, just as hundreds of volts coursing through wires in the walls once did. But like the wires in the walls, these hundreds of computers will come to be invisible to common awareness. People will simply use them unconsciously to accomplish everyday tasks.

Tabs are the smallest components of embodied virtuality. Because they are interconnected, tabs will expand on the usefulness of existing inch-scale computers, such as the pocket calculator and the pocket organizer. Tabs will also take on functions that no computer performs today. For example, computer scientists at PARC and other research laboratories around the world have begun working with active badges—clip-on computers roughly the size of an employee I.D. card, first developed by the Olivetti Cambridge research laboratory. These badges can identify themselves to receivers placed throughout a building, thus making it possible to keep track of the people or objects to which they are attached.

The software that symbolizes ease of use now has a symbol of its own.

Mac™ OS

As millions of Macintosh® personal computer users around the world will gladly confirm, no computing experience equals the feeling that comes with working on a Mac.™

That's because Macintosh is still the only personal computer designed from the beginning to work the way you do.

So when we designed this symbol to represent the Macintosh operating system, we included something you'd expect from the most human computer of all: a smile.

Today, you will find the Mac OS logo appearing on Macintosh-compatible hardware, software and peripherals everywhere.

It stands for simplicity and intuitive design. So you can be sure that the products you're buying are easy to set up and use.

It stands for compatibility with Macintosh and the thousands of business and scientific programs designed for the Mac.

And if you're looking for multimedia titles that are easy to work with, look for it as a sign of assurance that they'll run problem free.

Most important of all, the Mac OS logo stands for power. The kind of power you need to do more than ever. The power to be your best.®

Apple

© 1995 Apple Computer, Inc. All rights reserved. Apple, the Apple logo, Macintosh and "The power to be your best" are registered trademarks of Apple Computer, Inc. Mac and the Mac OS logo are trademarks of Apple Computer, Inc.

You got a Mac. And then a few more. You tied them together. People started sharing. Everything.

Now they're sharing their frustration. With you.

Time for a Workgroup Server.

Funny thing about the Macintosh* computer. With little fanfare and an idea known simply as "plug and play," it turned a whole lot of novice users into networkers.

And is it any surprise? All you have to do is plug a connector into your Macintosh, and you're networked. You can share printers. You can share modems. You can even share files with your peers.

This kind of "peer-to-peer" network is all some workgroups need.

Others, however, soon find they're ready to take the next step: a Workgroup Server dedicated to improving the ebb and flow of vital information.

How do you know if you're ready? Just ask yourself a few simple questions.

Do I have more than five computers tied together in a peer-to-peer network? Is one printer always backlogged while others sit idle? When people share files, are their systems slowing down? Is there confusion about who has the latest data and where it resides?

If you answered yes to any of these questions, it's definitely time to consider a Workgroup Server from Apple.

Workgroup Servers provide a centralized place for users to store and share files. Which means your shared files don't get lost, your systems don't slow down and users always have access to the latest information.

A server also gives you a place to put network applications and services—accounting programs, databases, e-mail and more—for everyone to use.

And that's just the start. With every Workgroup Server, you also get Dantz Retrospect Remote, so you can back up every system on the network—clients and servers—automatically. Plus Apple* RAID disk mirroring software, so you can rest easy about data reliability.

Want more still? Add FileWave software to your server, and you can easily distribute new programs and upgrades from the server. Or add Apple Remote Access software, and users can access your network from anywhere.

"OK," you say, "but how difficult is it to use a Workgroup Server?" It's as easy as using a Macintosh. All you have to do is connect the components, turn on the power, and you're ready to go.

Choosing the right server is just as easy. We have three models based on the new PowerPC™ RISC chip—the Workgroup Servers 6150, 8150 and 9150.

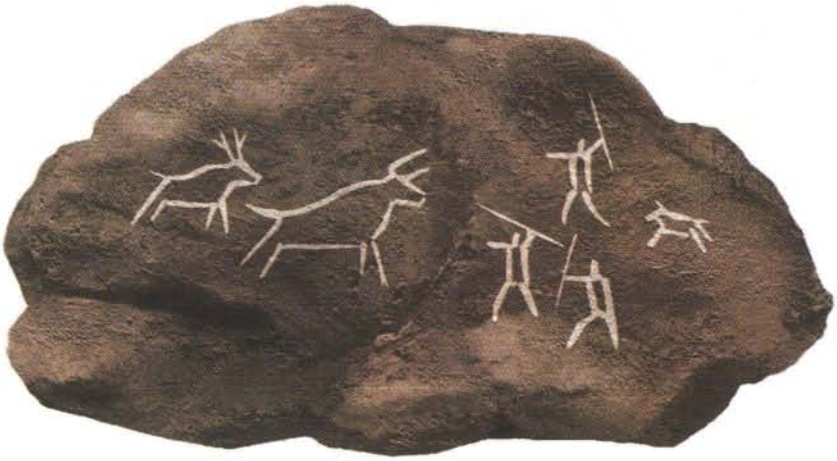
No matter which one you choose, you're backed by our expert phone support, 24 hours a day, 7 days a week.

And now, for a limited time, you'll get GraceLAN Server Manager software, a $395 retail value, at no extra charge.*

To get your free Workgroup Server brochure or for more information on our server trial program, simply pick up the phone and call us at 800-538-9696, ext. 430.** And unlock the power of your workgroup. The power to be your best.*

Apple

*Offer applies to Workgroup Servers 6150, 8150 and 9150 only. **Program ends March 15, 1995. Offer good while supplies last. ©1994 Apple Computer, Inc. All rights reserved. Apple, the Apple logo, Macintosh and "The power to be your best" are registered trademarks of Apple Computer, Inc. Mac is a trademark of Apple Computer, Inc. PowerPC is a trademark of International Business Machines Corporation, used under license therefrom. FileWave is a registered trademark of Wave Research, Inc. GraceLAN is a trademark of Technology Works, Inc. Retrospect Remote is a registered trademark of Dantz Development Corporation. Server Manager is a trademark and copyright of Santorini Consulting & Design, Inc.



The First Revolution In Communications.

Apple AT&T IBM SIEMENS

The Next.

versit

Apple, AT&T, IBM and Siemens are pleased to announce the formation of a historic communications initiative. It's called Versit™.

By making the convergence of computers and communications a reality, we'll be creating some incredible new ways for you to communicate and work with others – giving you more freedom than you ever dreamed possible.

To find out more, call us at 800-803-6240.* Or you can reach us via our Internet address: versit@cup.portal.com.

We look forward to hearing from you. Because you'll soon be hearing a lot from us.

*Outside the U.S., call 201-327-2803. ©1994 Apple, AT&T, IBM and Siemens. Versit and the Versit logo are trademarks of Apple Computer, Inc., AT&T Corp., International Business Machines Corp. and Siemens Rolm Communications, Inc. All other company logos are registered trademarks of their respective companies.

In our experimental embodied virtuality, doors open only to the right badge wearer, rooms greet people by name, telephone calls can be automatically forwarded to wherever the recipient may be, receptionists actually know where people are, computer terminals retrieve the preferences of whoever is sitting at them, and appointment diaries write themselves. The automatic diary shows how such a simple task as knowing where people are can yield complex dividends: meetings, for example, consist of several people spending time in the same room, and the subject of a meeting is most probably the files called up on that room's display screen while the people are there. No revolution in artificial intelligence is needed, merely computers embedded in the everyday world.

My colleague Roy Want has designed a tab incorporating a small display that can serve simultaneously as an active badge, calendar and diary. It will also act as an extension to computer screens: instead of shrinking a program window down to a small icon on the screen, for example, a user will be able to shrink the window onto a tab display. This will leave the screen free for information and also let people arrange their computer-based projects in the area around their terminals, much as they now arrange paper-based projects in piles on desks and tables. Carrying a project to a different office for discussion is as simple as gathering up its tabs; the associated programs and files can be called up on any terminal.

The next step up in size is the pad, something of a cross between a sheet of paper and current laptop and palmtop computers. Robert Krivacic of PARC has built a prototype pad that uses two microprocessors, a workstation-size display, a multibutton stylus and a radio network with enough communications bandwidth to support hundreds of devices per person per room.

Pads differ from conventional portable computers in one crucial way. Whereas portable computers go everywhere with their owners, the pad that must be carried from place to place is a failure. Pads are intended to be "scrap computers" (analogous to scrap paper) that can be grabbed and used anywhere; they have no individualized identity or importance.

One way to think of pads is as an antidote to windows. Windows were invented at PARC and popularized by Apple in the Macintosh as a way of fitting several different activities onto the small space of a computer screen at the same time. In 20 years computer screens have not grown much larger. Computer window systems are often said to be based on the desktop metaphor—but who would ever use a desk only nine inches high by 11 inches wide?

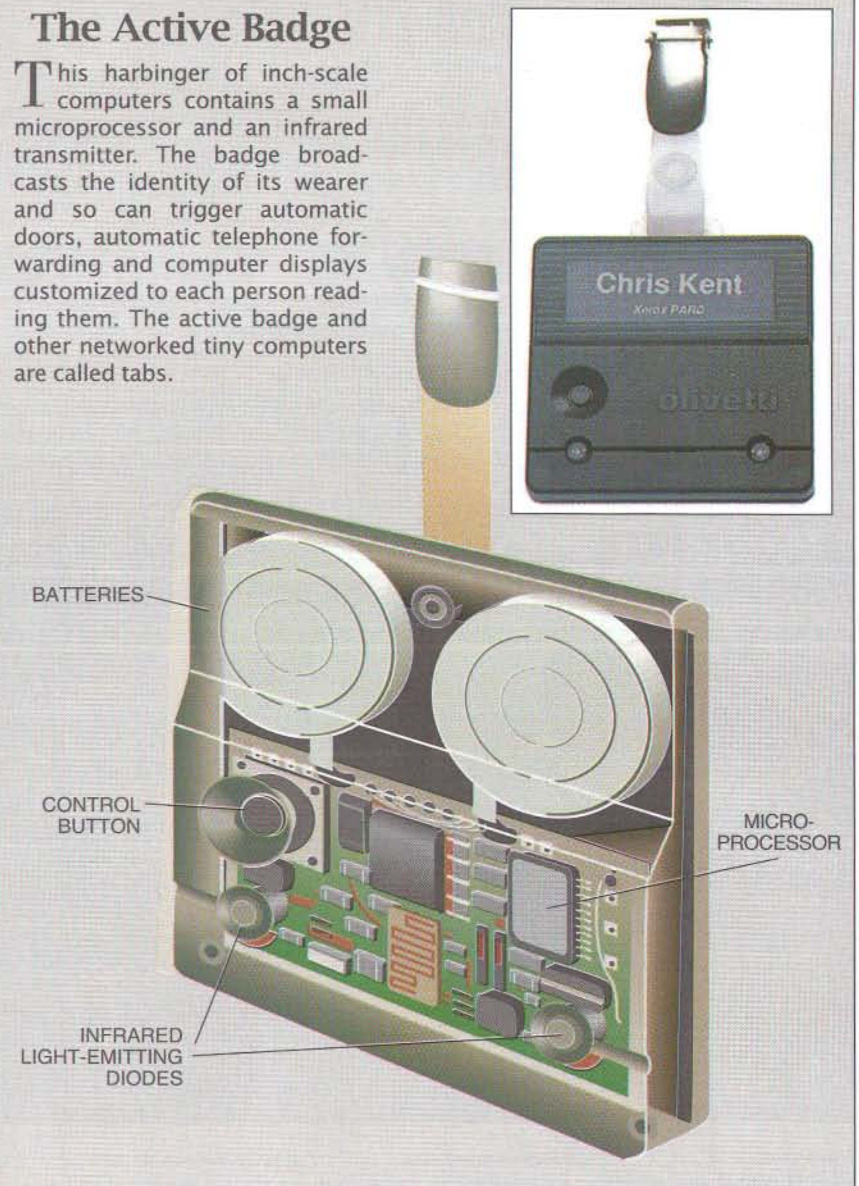
Pads, in contrast, use a real desk. Spread many electronic pads around on the desk, just as you spread out papers. Have many tasks in front of you, and use the pads as reminders. Go beyond the desk to drawers, shelves, coffee tables. Spread the many parts of the many tasks of the day out in front of you to fit both the task and the reach of your arms and eyes rather than to fit the limitations of glassblowing. Someday pads may even be as small and light as actual paper, but meanwhile they can fulfill many more of paper's functions than can computer screens.

Yard-size displays (boards) serve a number of purposes: in the home, video screens and bulletin boards; in the office, bulletin boards, white boards or flip charts. A board might also serve as an electronic bookcase from which one might download texts to a pad or tab. For the time being, however, the ability to pull out a book and place it comfortably on one's lap remains one of the many attractions of paper. Similar objections apply to using a board as a desktop; people will have to become accustomed to having pads and tabs on a desk as an adjunct to computer screens before taking embodied virtuality any further.

Prototype boards, built by Richard Bruce and Scott Elrod of PARC, are in use at several Xerox research laboratories. They measure about 40 by 60 inches and display 1,024 × 768 black-and-white pixels. To manipulate the

The Active Badge

This harbinger of inch-scale computers contains a small microprocessor and an infrared transmitter. The badge broadcasts the identity of its wearer and so can trigger automatic doors, automatic telephone forwarding and computer displays customized to each person reading them. The active badge and other networked tiny computers are called tabs.



display, users pick up a piece of wireless electronic "chalk" that can work either in contact with the surface or from a distance. Some researchers, using themselves and their colleagues as guinea pigs, can hold electronically mediated meetings or engage in other forms of collaboration around a live board. Others use the boards as testbeds for improved display hardware, new "chalk" and interactive software.

For both obvious and subtle reasons, the software that animates a large shared display and its electronic chalk is not the same as that for a workstation. Switching back and forth between chalk and keyboard may involve walking several steps, and so the act is qualitatively different from using a keyboard and mouse. In addition, body size is an issue. Not everyone can reach the top of the board, so a Macintosh-style menu bar might have to run across the bottom of the screen instead.

We have built enough live boards to permit casual use: they have been placed in ordinary conference rooms and open areas, and no one need sign up or give advance notice before using them. By building and using these boards, researchers start to experience and so understand a world in which computer interaction informally enhances every room. Live boards can usefully be shared across rooms as well as within them. In experiments instigated by Paul Dourish of EuroPARC and Sara Bly and Frank Halasz of PARC, groups at widely separated sites gathered around boards—each displaying the same image—and jointly composed pictures and drawings. They have even shared two boards across the Atlantic.

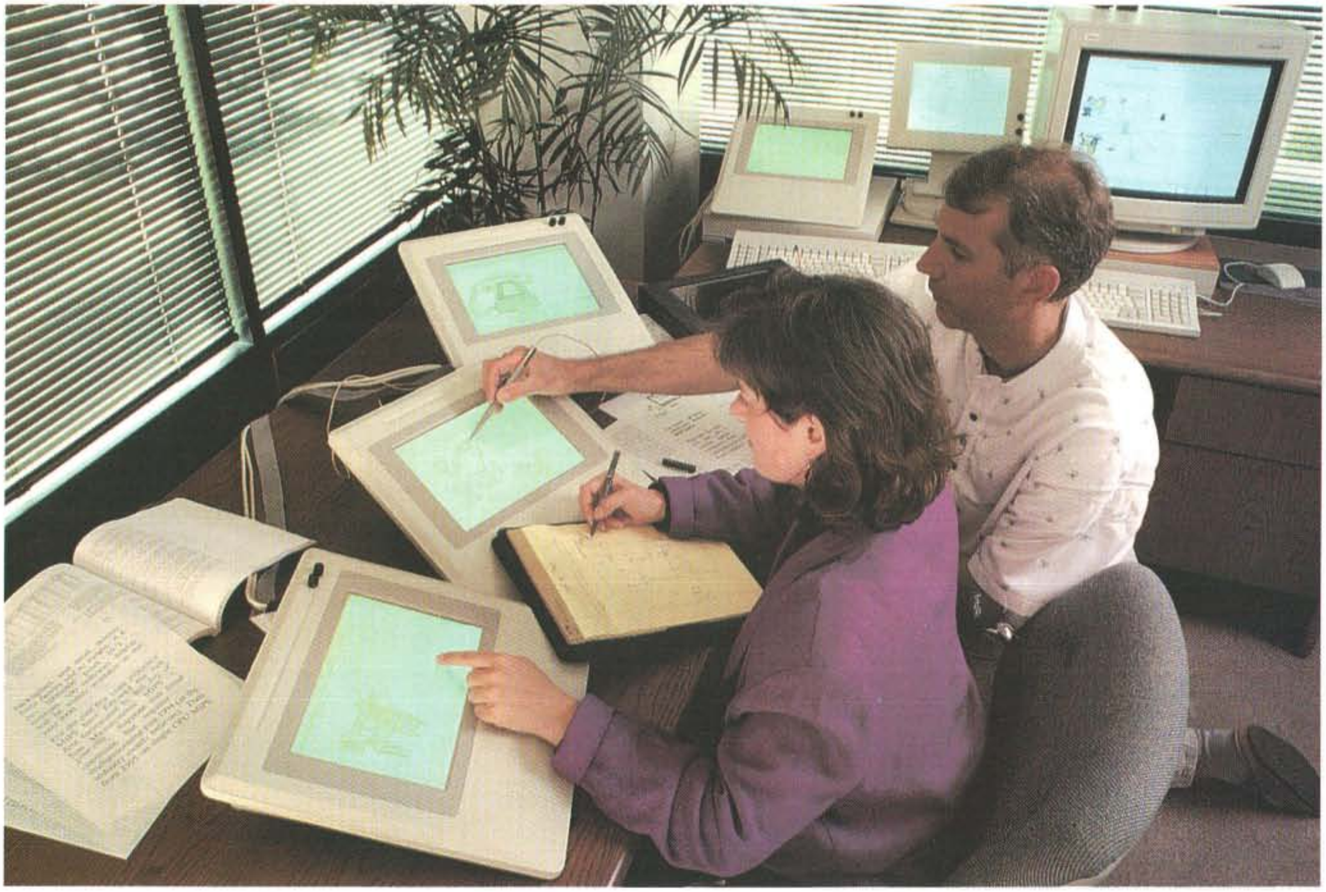
Live boards can also be used as bulletin boards. There is already too much text for people to read and comprehend all of it, and so Marvin Theimer and David Nichols of PARC have built a prototype system that attunes its public information to the people reading it. Their "scoreboard" requires little or no interaction from the user other than to look and to wear an active badge.

Prototype tabs, pads and boards are just the beginning of ubiquitous computing. The real power of the concept comes not from any one of these devices—it emerges from the interaction of all of them. The hundreds of processors and displays are not a "user interface" like a mouse and windows, just a pleasant and effective "place" to get things done.

What will be most pleasant and effective is that tabs can animate objects previously inert. They can beep to help locate mislaid papers, books or other items. File drawers can open and show the desired folder—no searching. Tabs in library catalogues can make active maps to any book and guide searchers to it, even if it is off the shelf, left on a table by the last reader.

In presentations, the size of text on overhead slides, the volume of the amplified voice, even the amount of ambient light, can be determined not by guesswork but by the desires of the listeners in the room at that moment. Software tools for tallying votes instantly and consensus checking are already available in electronic meeting rooms of some large corporations; tabs can make them widespread.

The technology required for ubiquitous computing comes in three parts: cheap, low-power computers that include equally convenient displays, software for ubiquitous applications and a network that ties them all together. Current trends suggest that



COMPUTER SCRATCHPADS augment the conventional screen in this office at the Xerox Palo Alto Research Center. Prototype pads are wired to conventional computers; thus far only a handful of wireless models have been built.

the first of these requirements will easily be met. Flat-panel displays containing 640 × 480 black-and-white pixels are now common. This is the standard size for PCs and is also about right for television. As long as laptop, palmtop and notebook computers continue to grow in popularity, display prices will fall, and resolution and quality will rise. By the end of the decade, a 1,000 × 800-pixel high-contrast display will be a fraction of a centimeter thick and weigh perhaps 100 grams. A small battery will provide several days of continuous use.

Larger displays are a somewhat different issue. If an interactive computer screen is to match a white board in usefulness, it must be viewable from arm's length as well as from across a room. For close viewing, the density of picture elements should be no worse than on a standard computer screen, about 80 per inch. Maintaining a density of 80 pixels per inch over an area several feet on a side implies displaying tens of millions of pixels. The biggest computer screen made today has only about one fourth that capacity. Such large displays will probably be expensive, but they should certainly be available.

The large display will require advanced microprocessors to feed it. Central-processing-unit speeds reached a million instructions per second in 1986 and continue to double each year. Some industry observers believe that this exponential growth in raw chip speed may begin to level off about 1994 but that other measures of performance, including power consumption and auxiliary functions, will still improve. The 100-gram flat-panel display, then, might be driven by a microprocessor that executes a billion operations per second and contains 16 megabytes of on-board memory along with sound, video and network interfaces. Such a processor would draw, on average, a few percent of the power required by the display.

Auxiliary storage devices will augment main memory capacity: conservative extrapolation of current technology suggests that removable hard disks (or nonvolatile memory chips) the size of a matchbook will store about 60 megabytes each. Larger disks containing several gigabytes of information will be standard, and terabyte storage—roughly the data content of the Library of Congress—will be common. Such enormous stores will not necessarily be filled to capacity with usable information. Abundant space will, however, allow radically different strategies of information management. A terabyte of disk storage will make deleting old files virtually unnecessary, for example.

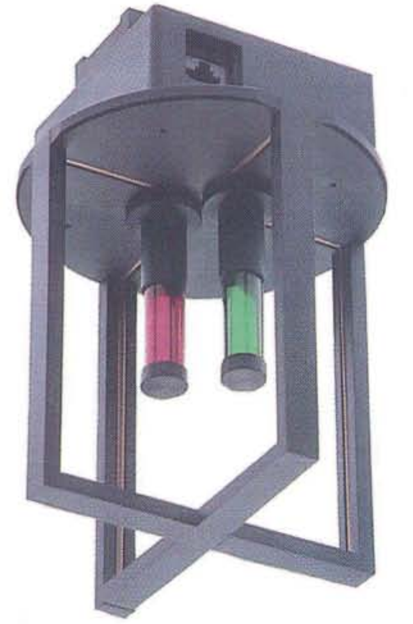
Although processors and displays should be capable of offering ubiquitous computing by the end of the decade, trends in software and network technology are more problematic. Current implementations of "distributed computing" simply make networked file servers, printers or other devices appear as if they were connected directly to each user's computer. This approach, however, does nothing to exploit the unique capabilities of physically dispersed computers and the information embodied in knowing where a particular device is located.

Computer operating systems and window-based display software will have to change substantially. The design of current operating systems, such as DOS and Unix, is based on the assumption that a computer's hardware and software configuration will not change substantially while it is running. This assumption is reasonable for conventional mainframes and personal computers, but it makes no sense in terms of ubiquitous computing. Pads, tabs and even boards may come and go at any time in any room, and it will certainly be impossible to shut down all the computers in a room to install new software in any one of them. (Indeed, it may be impossible to find all the computers in a room.)

One solution may be "micro-kernel" operating systems such as those developed by Rick Rashid of Carnegie Mellon University and A. S. Tanenbaum of Vrije University in Amsterdam. These experimental systems contain only the barest scaffolding of fixed computer code; software modules to perform specific functions can be readily added or removed. Future operating systems based on this principle could shrink and grow automatically to fit the changing needs of ubiquitous computation.

Current window display systems also are not ready to cope with ubiquitous computing. They typically assume that a particular computer will display all the information for a single application. Although the X Window System and Windows 3.0, for example, can cope with multiple screens, they do not do well with applications that start out on one screen and move to another, much less those that peregrinate from computer to computer or room to room.

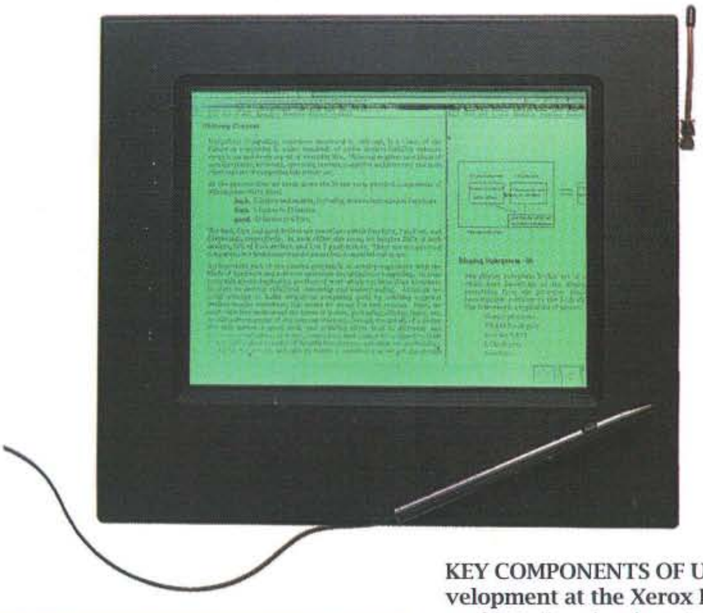
Solutions to this problem are in their infancy. Certainly no existing display system can perform well while working with the full diversity of input and output forms required by embodied virtuality. Making pads, tabs and boards work together seamlessly will require changes in the kinds of protocols by which applications programs and their displayed windows communicate.

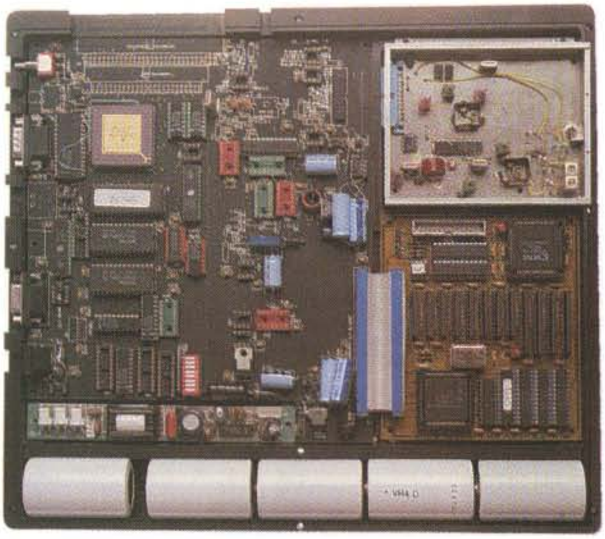


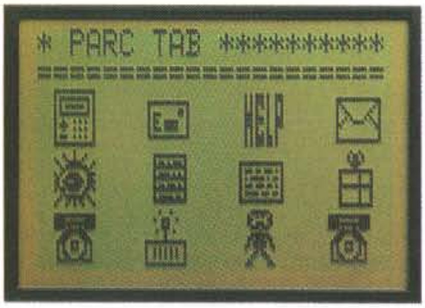
RADIO TRANSCEIVER links pads and other movable computer devices to the wired network. This unit, intended to be mounted on the ceiling, contains antennas in its crossed arms and two light-emitting diodes to signal its status.

The network that will connect ubiquitous hardware and software poses further challenges. Data transmission rates for both wired and wireless networks are increasing rapidly. Access to gigabit-per-second wired nets is already possible, although expensive, and will become progressively cheaper. (Gigabit networks will seldom devote all of their bandwidth to a single data stream; instead they will allow enormous numbers of lower-speed transmissions to proceed simultaneously.) Small wireless networks, based on digital cellular telephone principles, currently offer data rates between two and 10 megabits per second over a range of a few hundred meters. Low-power wireless networks capable of transmitting 250,000 bits per second to each station will eventually be available commercially.

Yet the problem of transparently linking wired and wireless networks resists solution. Although some stopgap methods have been developed, engineers must develop new communications protocols that explicitly recognize the concept of machines that move in physical space. Furthermore, the number of channels envisioned in most wireless network schemes is still







KEY COMPONENTS OF UBIQUITOUS COMPUTING are the pads and tabs under development at the Xerox Palo Alto Research Center. The page-size pad (*top, exterior and interior views*) contains two microprocessors, four million bytes of random-access memory, a high-speed radio link, a high-resolution pen interface and a black-and-white display that is 1,024 by 768 pixels. Because it uses standard window system software, the pad can communicate with most workstations. The much smaller tab (*at left*), $2^3/4$ by $3^1/4$ inches, has three control buttons, a pen interface, audio and an infrared link for communicating throughout a room. The author believes future homes and offices will contain hundreds of these tiny computers.

very small, and the range large (50 to 100 meters), so that the total number of mobile devices is severely limited. The ability of such a system to support hundreds of machines in every room is out of the question. Single-room networks based on infrared or newer electromagnetic technologies have enough channel capacity for ubiquitous computers, but they can work only indoors.

Present technologies would require a mobile device to have three different network connections: tiny-range wireless, long-range wireless and very high speed wired. A single kind of network connection that can somehow serve all three functions has yet to be invented.

Neither an explication of the principles of ubiquitous computing nor a list of the technologies involved really gives a sense of what it would be like to live in a world full of invisible widgets. Extrapolating from today's rudimentary fragments of embodied virtuality is like trying to predict the publication of *Finnegans Wake* shortly after having inscribed the first clay tablets. Nevertheless, the effort is probably worthwhile:

Sal awakens; she smells coffee. A few minutes ago her alarm clock, alerted by her restless rolling before waking, had quietly asked, "Coffee?" and she had mumbled, "Yes." "Yes" and "no" are the only words it knows.

Sal looks out her windows at her neighborhood. Sunlight and a fence are visible through one, and through others she sees electronic trails that have been kept for her of neighbors coming and going during the early morning. Privacy conventions and practical data rates prevent displaying video footage, but time markers and electronic tracks on the neighborhood map let Sal feel cozy in her street.

Glancing at the windows to her kids' rooms, she can see that they got up 15 and 20 minutes ago and are already in the kitchen. Noticing that she is up, they start making more noise.

At breakfast Sal reads the news. She still prefers the paper form, as do most people. She spots an interesting quote from a columnist in the business section. She wipes her pen over the newspaper's name, date, section and page number and then circles the quote. The pen sends a message to the paper, which transmits the quote to her office.

Electronic mail arrives from the company that made her garage door opener. She had lost the instruction manual and asked them for help. They have sent her a new manual and also something unexpected—a way to find the old one. According to the note, she can press a code into the opener and the missing manual will find itself. In the garage, she tracks a beeping noise to where the oil-stained manual had fallen behind some boxes. Sure enough, there is the tiny tab the manufacturer had affixed in the cover to try to avoid E-mail requests like her own.

On the way to work Sal glances in the foreview mirror to check the traffic. She spots a slowdown ahead and also notices on a side street the telltale green in the foreview of a food shop, and a new one at that. She decides to take the next exit and get a cup of coffee while avoiding the jam.

Once Sal arrives at work, the foreview helps her find a parking spot quickly. As she walks into the building, the machines in her office prepare to log her in but do not complete the sequence until she actually enters her office. On her way, she stops by the offices of four or five colleagues to exchange greetings and news.

Sal glances out her windows: a gray day in Silicon Valley, 75 percent humidity and 40 percent chance of afternoon showers; meanwhile it has been a quiet morning at the East Coast office. Usually the activity indicator shows at least one spontaneous, urgent meeting by now. She chooses not to shift the window on the home office back three hours—too much chance of being caught by surprise. But she knows others who do, usually people who never get a call from the East but just want to feel involved.

The telltale by the door that Sal programmed her first day on the job is blinking: fresh coffee. She heads for the coffee machine.

Coming back to her office, Sal picks up a tab and "waves" it to her friend Joe in the design group, with whom she has a joint assignment. They are sharing a virtual office for a few weeks. The sharing can take many forms—in this case, the two have given each other access to their location detectors and to each other's screen contents and location. Sal chooses to keep miniature versions of all Joe's tabs and pads in view and three-dimensionally correct in a little suite of tabs in the back corner of her desk. She can't see what anything says, but she feels more in touch with his work when noticing the displays change out of the corner of her eye, and she can easily enlarge anything if necessary.

A blank tab on Sal's desk beeps and displays the word "Joe" on it. She picks it up and gestures with it toward her live board. Joe wants to discuss a document with her, and now it shows up on the wall as she hears Joe's voice:

"I've been wrestling with this third paragraph all morning, and it still has the wrong tone. Would you mind reading it?"

Sitting back and reading the paragraph, Sal wants to point to a word. She gestures again with the "Joe" tab onto a nearby pad and then uses the stylus to circle the word she wants:

"I think it's this term 'ubiquitous.' It's just not in common enough use and makes the whole passage sound a little formal. Can we rephrase the sentence to get rid of it?"

"I'll try that. Say, by the way, Sal, did you ever hear from Mary Hausdorf?"

"No. Who's that?"

"You remember. She was at the meeting last week. She told me she was going to get in touch with you."

Sal doesn't remember Mary, but she does vaguely remember the meeting. She quickly starts a search for meetings held during the past two weeks with more than six people not previously in meetings with her and finds the one. The attendees' names pop up, and she sees Mary.

As is common in meetings, Mary made some biographical information about herself available to the other attendees, and Sal sees some common background. She'll just send Mary a note and see what's up. Sal is glad Mary did not make the biography available only during the time of the meeting, as many people do....

In addition to showing some of the ways that computers can enter invisibly into people's lives, this scenario points up some of the social issues that embodied virtuality will engender. Perhaps key among them is privacy: hundreds of computers in every room, all capable of sensing people near them and linked by high-speed networks, have the potential to make totalitarianism up to now seem like sheerest anarchy. Just as a workstation on a local-area network can be programmed to intercept messages meant for others, a single rogue tab in a room could potentially record everything that happened there.

Even today the active badges and self-writing appointment diaries that offer all kinds of convenience could be a source of real harm in the wrong hands. Not only corporate superiors or underlings but also overzealous government officials and even marketing firms could make unpleasant use of the same information that makes invisible computers so convenient.

Fortunately, cryptographic techniques already exist to secure messages from one ubiquitous computer to another and to safeguard private information stored in networked systems. If designed into systems from the outset, these techniques can ensure that private data do not become public. A well-implemented version of ubiquitous computing could even afford better privacy protection than exists today.

By pushing computers into the background, embodied virtuality will make individuals more aware of the people on the other ends of their computer links. This development may reverse the unhealthy centripetal forces that conventional personal computers have introduced into life and the workplace.

Even today, people holed up in windowless offices before glowing computer screens may not see their fellows for the better part of each day. And in virtual reality, the outside world and all its inhabitants effectively cease to exist. Ubiquitous computers, in contrast, reside in the human world and pose no barrier to personal interactions. If anything, the transparent connections that they offer between different locations and times may tend to bring communities closer together.

My colleagues and I at PARC believe that what we call ubiquitous computing will gradually emerge as the dominant mode of computer access over the next 20 years. Like the personal computer, ubiquitous computing will produce nothing fundamentally new, but by making everything faster and easier to do, with less strain and fewer mental gymnastics, it will transform what is apparently possible. Desktop publishing, for example, is essentially no different from computer typesetting, which dates back to the mid-1960s. But ease of use makes an enormous difference.

When almost every object either contains a computer or can have a tab attached to it, obtaining information will be trivial: "Who made that dress? Are there any more in the store? What was the name of the designer of that suit I liked last week?" The computing environment knows the suit you looked at for a long time last week because it knows both of your locations, and it can retroactively find the designer's name even though that information did not interest you at the time.

Sociologically, ubiquitous computing may mean the decline of the computer addict. In the 1910s and 1920s many people "hacked" on crystal sets to take advantage of the new high-tech world of radio. Now crystal-and-cat's-whisker receivers are rare because high-quality radios are ubiquitous. In addition, embodied virtuality will bring computers to the presidents of industries and countries for nearly the first time. Computer access will penetrate all groups in society.

Most important, ubiquitous computers will help overcome the problem of information overload. There is more information available at our fingertips during a walk in the woods than in any computer system, yet people find a walk among trees relaxing and computers frustrating. Machines that fit the human environment instead of forcing humans to enter theirs will make using a computer as refreshing as taking a walk in the woods.

FURTHER READING

THE TACIT DIMENSION. Michael Polanyi. Doubleday & Company, 1966.

TOWARD PORTABLE IDEAS. Mark Stefik and John Seely Brown in *Technological Support for Work Group Collaboration.* Edited by Margrethe H. Olson. Lawrence Erlbaum Associates, 1989.

RECENT DEVELOPMENTS IN OPERATING SYSTEMS. Special issue of *Computer (IEEE Computer Society),* Vol. 23, No. 5; May 1990.

ACTIVITY-BASED INFORMATION RETRIEVAL: TECHNOLOGY IN SUPPORT OF HUMAN MEMORY. Mik Lamming and William Newman. Available as Rank Xerox EuroPARC Technical Report 91-03; February 1991.

LCDS AND BEYOND. Nick Baran in *Byte,* Vol. 16, No. 2, pages 229–236; February 1991.

A TALK WITH INTEL. Kenneth M. Sheldon in *Byte,* Vol. 16, No. 4, pages 131–140; April 1991.

The Future of the Transistor

As it has grown smaller and cheaper, engineers have scoffed at theoretical barriers to its progress—so far

by Robert W. Keyes

I am writing this article on a computer that contains some 10 million transistors, an astounding number of manufactured items for one person to own. Yet they cost less than the hard disk, the keyboard, the display and the cabinet. Ten million staples, in contrast, would cost about as much as the entire computer. Transistors have become this cheap because during the past 40 years engineers have learned to etch ever more of them on a single wafer of silicon. The cost of a given manufacturing step can thus be spread over a growing number of units.

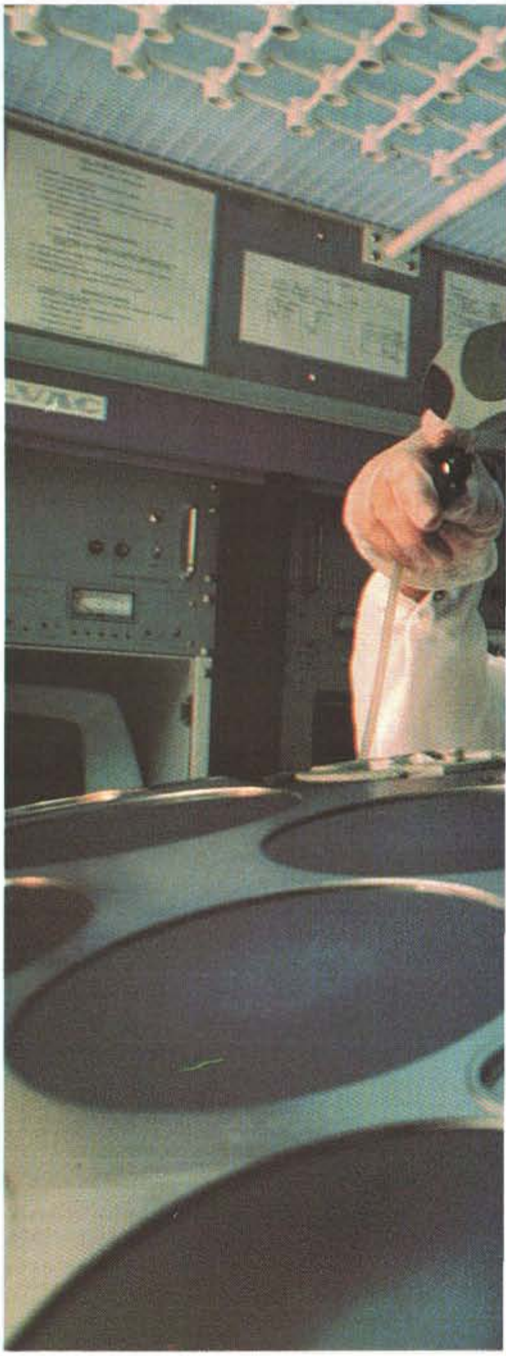
How much longer can this trend continue? Scholars and industry experts have declared many times in the past that some physical limit exists beyond which miniaturization could not go. An equal number of times they have been confounded by the facts. No such limit can be discerned in the quantity of transistors that can be fabricated on silicon, which has proceeded through eight orders of magnitude in the 46 years since the transistor was invented [*see box on page 94*].

I do not have a definitive answer to the question of limits. I do, however, have some thoughts on how the future of solid-state electronics will develop and what science is needed to support continuing progress.

Several kinds of physical limitations might emerge as the size of the transistor continues to shrink. The task of connecting minute elements to one another might, for example, become impossible. Declining circuit size also means that researchers must cope with ever stronger electric fields, which can affect the movement of electrons in several ways. In the not too distant future the transistor may span only hundreds of angstroms. At that point, the presence or absence of individual atoms, as well as their behavior, will become significant. Diminishing size leads to increasing density of transistors on a chip, which raises the amount of waste heat that is thrown off. Today's chips shed about 10 times as much heat as does a cooking surface of comparable size, a flux that can rise at least 10-fold without exceeding the cooling capacity of known designs. As the size of circuit elements drops below the wavelength of usable forms of radiation, existing manufacturing methods may reach their limits.

ROBERT W. KEYES recently retired from the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y., and works as a consultant. His interests have focused on semiconductor physics and devices and on the physics of information-processing systems, subjects on which he has written and lectured widely; he has received eight patents. Keyes studied at the University of Chicago, where he earned a doctorate in physics. He has participated in the programs of the National Research Council, the American Physical Society and the Institute of Electrical and Electronic Engineers.

To see how such problems might arise and how they can be addressed, it is useful to review the operation of the field-effect transistor, the workhorse of modern data processing. Digital computers operate by manipulating statements made in a binary code, which consists of ones and zeroes. A field-effect transistor is operated so that, like a relay, it is switched only "on" or "off." The device therefore represents exactly one binary unit of information: a bit. In a large-scale system, input signals control transistors, establishing connections that produce signals on output wires. The wires carry the signals to other switches that produce outputs, which are again sent on to another stage. The connections within the computer and the way in which the input signals determine an output together represent a logical statement. A series of such statements, in turn, determines a word in a document or an entry in a spreadsheet.



MINIATURIZATION has made transistors cheaper than staples by spreading man-

Reprinted from the June 1993 issue

The field-effect transistor contains a channel that interacts with three electrodes: a source, which supplies electrons to the channel; a drain, which receives them at the other side; and a gate, which influences the conductivity of the channel [*see illustration on next page*]. Each part contains different impurity atoms, or dopants, which modify the electrical properties of the silicon.

The gate switches the transistor on when a positive voltage applied to it attracts electrons to the interface between the semiconductor and the gate insulator. These electrons then establish a connection between the source and drain electrodes that allows current to be passed between them. At this point, the transistor is "on." The connection persists for as long as the positive charge remains on the gate. An incoming signal is applied to the gate and thus determines whether the connection between source and drain is established. If a connection results, the output is connected to the ground potential, one of the standard digital voltages. If no connection results, the output is connected through the resistor to the positive power supply, the other standard digital voltage.

Circuits of transistors must be oblivious to the operations of neighboring arrays. Existing concepts of insulation, impedance and other basic electrical properties of semiconductors and their connections should work well enough, for designers' purposes, in the next generation of devices. It is only when conducting areas approach to within about 100 angstroms of one another that

ufacturing costs over millions of devices on each of the hundreds of chips on a single wafer. This worker holds a nearly completed wafer. Its components will be connected by the condensation of metal in a vacuum chamber (*foreground*).

quantum effects, such as electron tunneling, threaten to create problems. In laboratory settings, researchers are already at the brink of this limit, at about 30 angstroms; in commercial devices, perhaps a decade remains before that limit is reached.

Another challenge is the strengthening of the electric field that inevitably accompanies miniaturization. This tendency constrains the design of semiconductor devices by setting up a basic conflict. Fields must continually get stronger as electron pathways shrink, yet voltages must remain above the minimum needed to overwhelm the thermal energy of electrons. In silicon at normal operating temperatures, the thermal voltage is 0.026 electron volt. Therefore, whenever a semiconductor is switched so as to prevent the passage of electrons, its electrical barrier must be changed by a factor several times as large. One can minimize the thermal problem by chilling the chip (an expensive proposition).

Even cooling cannot end the problem of the electric field. Signals must still have the minimum voltage that is characteristic of a semiconductor junction. In silicon this electrical barrier ranges between half a volt and a volt, depending on the degree of doping. That small voltage, applied over a very short distance, suffices to create an immensely strong electric field. As electrons move through such a field, they may gain so much energy that they stimulate the creation of electron-hole pairs, which are themselves accelerated. The resulting chain reaction can cause an avalanche of rising current, thereby disrupting the circuit. An avalanche begins when fields exceed about 500,000 volts per centimeter, compared with about 400,000 volts in today's chips.

Workers resort to a variety of tricks to mitigate the effects of strong electric fields. They have designed field-effect transistors, for example, in which the field can be moved to a place where it does not disrupt other electronic functions. This strategem is just one of many, all of which entail trade-offs with other desired characteristics, such as simplicity of design, ease of manufacture, reliability and long working life.

Miniaturization also increases the heat given off by each square centimeter of silicon. The reason is purely geometric: electrical pathways, and their associated energy losses, shrink in one dimension, whereas chip area shrinks in two. That relation means that as circuits get smaller, unit heat generation falls, albeit more slowly than does the number of units per square centimeter.

Devices already pour out as much as 30 watts per square centimeter, a radiance that one would expect of a material heated to about 1,200 degrees Celsius. Of course, the chips cannot be allowed to reach such temperatures, and so cooling systems remove heat as fast as it is produced. We have not even come close to exhausting the existing cooling technologies, which can siphon off energy many times faster than is now required.

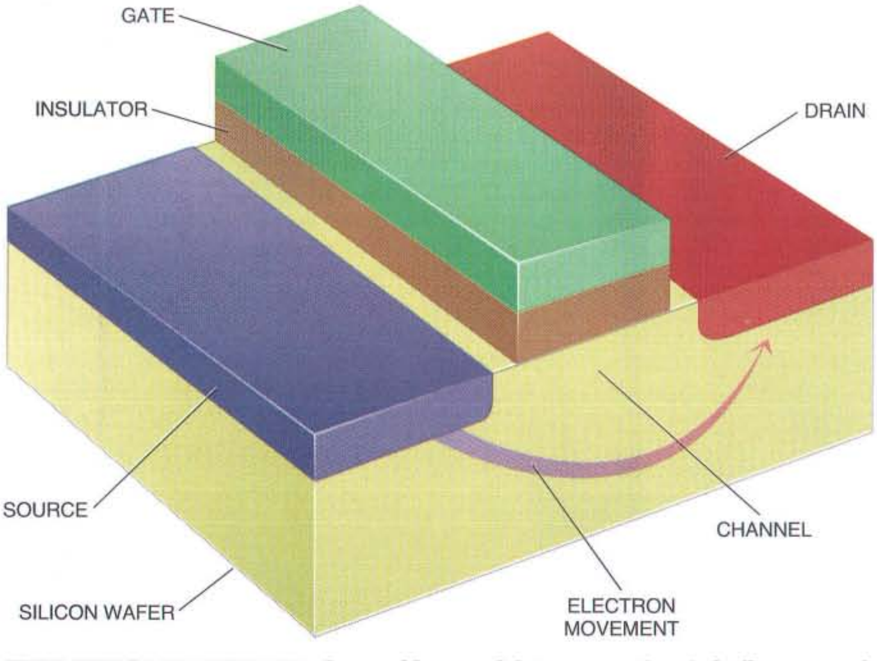
The exigencies of manufacturing impose constraints on the performance of electronic devices that might not be apparent from a purely theoretical discussion. Low-cost manufacturing leads, perforce, to small differences among the devices that are made on each wafer, as well as among those fabricated on different wafers. This variability cannot be banished—it is inherent in the way solid-state devices are made.

A semiconducting material, such as silicon, is made into a transistor in an integrated process involving many steps. Templates, called masks, are applied to the silicon in order to expose desired areas. Next, various operations involving chemical diffusion, radiation, doping, sputtering or the deposition of metal act on these areas, sometimes by constructing device features, other times by erecting scaffolding to be used in succeeding steps and then torn down. Meanwhile other devices—resistors, capacitors and conductors—are being built to connect the transistors.

Variations intrude at every step. For example, perfect focusing of the source of radiation over a large wafer is hard to achieve. The temperature of the wafer may vary slightly from one place to another during processing steps, causing a difference in the rate of chemical reactions. The mixing of gases in a reaction chamber may not be perfect. For many reasons, the properties of devices on a given wafer and between those on different wafers are not identical. Indeed, some devices on a wafer may be no good at all; the proportion of such irremediable errors places a practical limit on the size of an integrated circuit.

A certain amount of fuzziness is inherent in optical exposures. The light used in photolithography is diffracted as it passes through the holes in the template. Such diffraction can be minimized by resorting to shorter wavelengths.

When photolithographic fabrication was invented in the early 1970s, white light was used. Workers later switched to monochromatic laser light, moving up the spectrum until, in the mid-1980s, they reached the ultraviolet wavelengths. Now the most advanced commercial chips are etched by deep ultraviolet light, a difficult operation because it is hard to devise lasers with output in that range. The next generation of devices may require x-rays. In-



FIELD-EFFECT TRANSISTOR, the workhorse of data processing, is built as a sandwich of variously doped silicon layers. It contains a channel, a source, a drain and an insulated gate. When a positive voltage is applied to the gate, electrons move near the insulation, establishing a connection underneath it that allows current to pass from source to drain, switching the transistor on.

deed, each generation of circuitry requires manufacturing equipment of unprecedented expense.

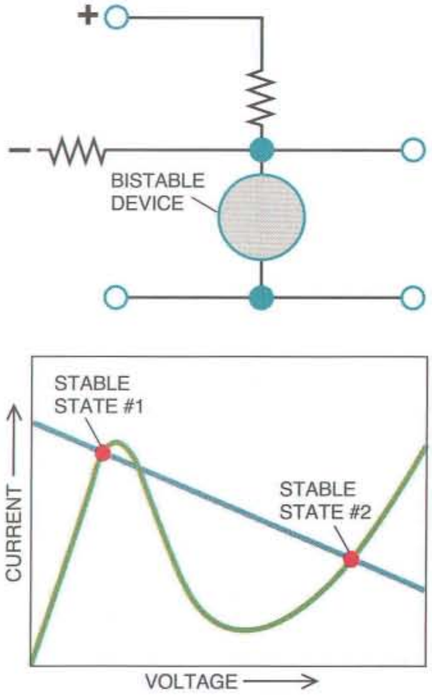
Other problems also add to the cost of making a chip. The mechanical controls that position wafers must become more precise. The "clean rooms" and chambers must become ever cleaner to ward off the ever smaller motes that can destroy a circuit. Quality-control procedures must become more elaborate as the number of possible defects on a chip increases.

Miniaturization may at first glance appear to involve manipulating just the width and breadth of a device, but depth matters as well. Sometimes the third dimension can be a valuable resource, as when engineers sink capacitors edgewise into a chip to conserve space on the surface. At other times, the third dimension can constrain design. Chip designers must worry about the aspect ratio, that is, the relation of depth to surface area. Their device resembles a sandwich (sometimes even a Dagwood sandwich). If the layers are multiplied excessively, etching will tend to undercut the structure before it can be completed. Such constraints compel workers to achieve tighter control over the thickness of films and over the penetration of impurities into the semiconductor. At present levels of control, about 20 layering operations—called mask steps—are feasible.

The formulas that are used to design large devices cannot be used for the tiny transistors now being made in laboratories. Designers need to account for exotic new phenomena that appear in such extremely small devices. Because the effects cannot be accurately treated by purely analytic methods, the designers have recourse to computer models that simulate the motion of electrons in a device.

A computer follows a single electron through a device, keeping track of its position as time is increased in small steps. Physical theory and experimental information are used to calculate the probability of the various events that are possible. The computer uses a table for the probabilities, stored in its memory, and a random number generator to simulate the occurrence of these events. For example, an electron is accelerated by an electric field, and the direction of its motion might be changed by a collision with an impurity. Adding the results of thousands of electrons modeled in this fashion gives a picture of the response of the device.

Consider the seemingly trivial question of how to represent the motion of an electron within an electric field. When path lengths were comparatively long, an electron quickly accelerated to the point at which collisions robbed it of energy as fast as the field supplied new energy. The particle therefore spent most of its time at a constant velocity, which can be modeled by a simple, linear equation. But when path lengths became shorter, the electron no longer had time to reach a stable velocity. The particles now accelerate all the time, and the equations must account for that complication.



BISTABLE CIRCUIT does the transistor's job by exploiting nonlinear effects. A device such as a tunnel diode is placed at the junction of two main electrodes and a minor one (*top*). If the minor electrode injects some extra current, the circuit will move from one stable state to the other (*bottom*). Such devices are impractical because they cannot tolerate much variation in signal strength.

If such difficulties can arise in modeling a well-understood phenomenon, what lies ahead as designers probe the murky physics of the ultrasmall? Simulations can be no better than the models that physicists make of events that happen in small spaces during short periods. To refine these models, researchers need to carry out experiments on femtosecond time scales.

More knowledge of solid-state physics is required, because as chips grow more complex they require more fabrication steps, and each step can influence the next. For instance, when doping atoms are introduced into a crystal, they tend to attract, repel or otherwise affect the motion of other dopants. Such effects of dopants on other dopants are not well understood; further experiments and theoretical investigations are therefore needed. Chemical reactions that take place on the surface of a silicon crystal demand a supply of silicon atoms, a kind of fluid flow within the solid lattice; does such motion carry other constituents along with it? These questions did not exercise designers of earlier generations of chips, because the transistors were then large enough to swamp such ultramicroscopic tendencies.

The prospect of roadblocks aside, the transistor has only itself to blame for speculation about alternative technologies. Its extraordinary success in the 1950s stimulated an explosive development of solid-state physics. In the course of the work, investigators discovered many other phenomena, which in turn suggested a host of ideas for electronic devices. Several of these lines of research produced respectable bodies of new engineering knowledge but none that led to anything capable of even finding a niche in information processing.

Some workers have argued that the transistor owes its preeminence to being first off the block. Because of that head start, semiconductors have been the center of research, a position that guarantees them a margin of technological superiority that no rival can match. Yet I believe the transistor has intrinsic virtues that, by themselves, could probably preserve its dominant role for years to come.

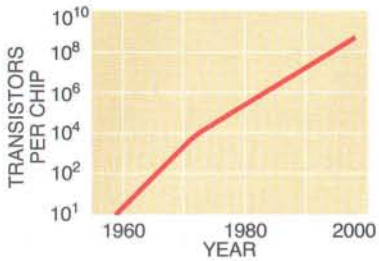
I participated, as a minor player, in some of the efforts to build alternative switches, the repeated failures of which made me wonder what was missing. Of course, quite a few new fabrication methods had to be developed to implement a novel device concept. But even though these could be mastered, it was difficult to get a large collection of components to work together.

What gave the transistor its initial, sudden success? One difference stood out: the transistor, like the vacuum tube before it, has large gain. That is, it is capable of vastly amplifying signals of the kind processed in existing circuits, so that a small variation in input can produce a large variation in output. Gain makes it possible to preserve the integrity of a signal as it passes through many switches.

Rivals to the transistor may have been equally easy to miniaturize, but they exhibited far less gain. Take, for instance, bistable devices [*see illustration on this page*], which perform log-

The Shrinking Transistor

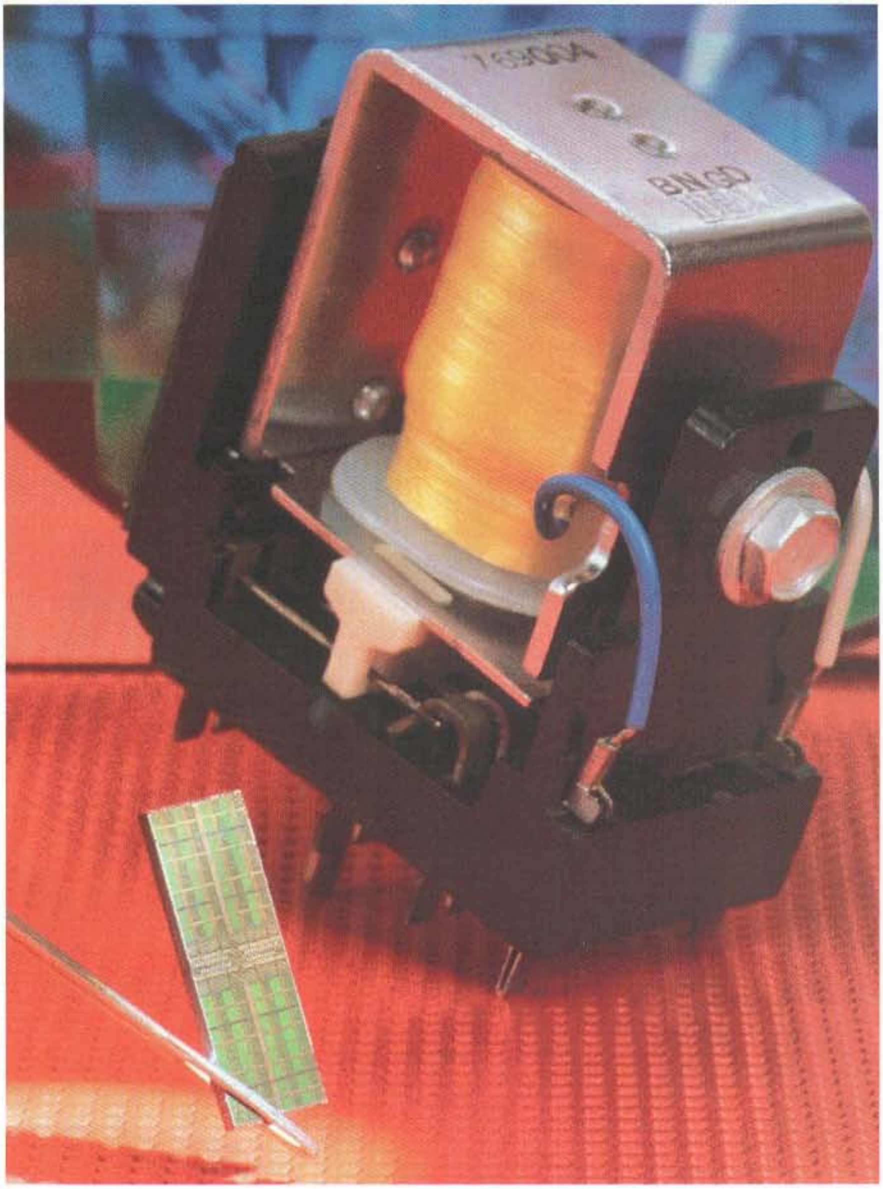
Miniaturization is manifest in this comparison between an electromechanical switch, circa 1957, and a recent chip containing 16 million bits of memory (*right*). Progress appears in these snapshots (*counterclockwise*): Bell Laboratories' first transistor (1948); canned transistors (1958); salt-size transistors (1964); 2,000-bit chip (1973); boards with 185,000 circuits and 2.3 megabits of memory apiece (1985); and a 64-megabit memory chip (1992).



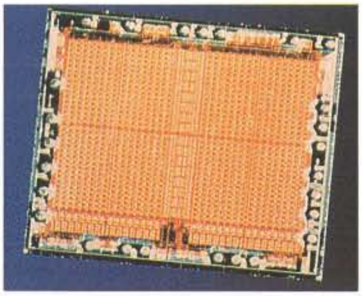
The first transistor

Early commercial transistors

Salt-size transistors

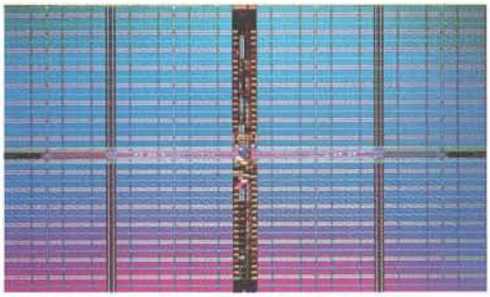


Switches, now and then



Early integrated circuit

Circuit assembly



Contemporary memory chip

ic functions by moving between two stable states that are separated by an unstable transition. Researchers have produced such a transition by designing circuits having a range of values in which current declines as voltage increases. Any slight disturbance, such as that obtained by injecting extra current through the device, will switch the circuit between its two stable states. Because this slight input can bring about large changes in the current and voltages, there is a sense in which gain is achieved.

Yet the gain is far less useful than that provided by a transistor because it operates within narrow tolerances. A bistable switch thus performs deceptively well in the laboratory, where one can fine-tune the circuit so it stays near the crossover point. A collection of such switches, however, does not lend itself to such painstaking adjustments. Because not all the circuits will work, no complex device can be based on their operation. Negative resistance therefore plays no role in practical data processing.

IMMENSE AND DENSE: this active-matrix liquid-crystal panel shows that today's electronic structures can achieve great complexity over large areas. Each liquid-crystal pixel is controlled by its own transistor, providing extraordinary resolution.

The same difficulty has plagued the development of nonlinear optical devices, in which the intensity of optical beams replaces the currents and voltages of electrical circuits. Here, too, the operation depends on fine-tuning the system so that a small input will upset a delicate balance. (Such switches have occasionally been termed "optical transistors," a label that misconstrues the principles of transistor action.) Optical switches face a problem even more fundamental. Light, unlike electricity, hardly interacts with light, yet the interaction of signals is essential for logic functions. Optical signals must therefore be converted into electrical ones in a semiconductor. The voltage thus produced changes the optical response of another material, thereby modulating a beam of light.

Another proposed switch, sometimes called a quantum interference device, depends on the interference of waves. In the most familiar case, that of electromagnetic radiation, or light, one wave is divided into two components. The components begin oscillating in phase, that is, their peaks and troughs vibrate in tandem. If the components follow routes of different lengths before reuniting, the phase relation between their waveforms will be changed. Consequently, the peaks and troughs either cancel or reinforce one another, producing a pattern of bright and dark fringes. The displacement of the fringes measures the relative phase of the system.

Electrons also possess a wave nature and can be made to interfere. If the two components of a wave move at equal speeds over similar paths to a rendezvous, they will reconstitute the original wave; if they move at different speeds, they will interfere. One can manipulate the velocity of one wave by applying a tiny electric field to its pathway. The correct field strength will cause the waves to cancel so that no current can flow through the device.

At first sight, this action duplicates that of a field-effect transistor, which uses an electric field to control a current through a semiconductor. In an interference device, however, conditions must be just right: if the applied voltage is too high or too low, there will be some current. This sensitivity means that an interference device will not restore the binary nature of a degraded input signal but will instead add its own measure of noise. Data passing from one such device to another will quickly degenerate into nothingness.

The lack of real rivals means that the future of digital electronics must be sought in the transistor. The search begins anew with each voyage into a smaller scale or a different material. The latest reexamination was occasioned by the introduction of new semiconductor materials, such as gallium arsenide and related compounds, several of which may even be incorporated into a single device. These combinations may be used to produce what are called heterojunctions, in which crystalline lattices of different energy gaps meet. Lattices may mesh imperfectly, creating atomic-scale defects, or they may stretch to fit one another, in the process creating an elastic strain. Either defects or strain can produce electrical side effects.

These combinations complicate the physics but at the same time provide a variable that may be useful in surmounting the many design problems that miniaturization creates. For instance, the dopants that supply electrons to a semiconductor also slow the electrons. To reduce this slowing effect, one can alternate layers of two semiconductors in which electrons have differing energies. The dopants are placed in the high-energy semiconductor, but the electrons they donate immediately fall into the lower-energy layers, far from the impurities.

What, one may ask, would one want with a technology that can etch a million transistors into a grain of sand or put a supercomputer in a shirt pocket? The answer goes beyond computational power to the things such power can buy in the emerging information economy. It has only recently been taken for granted that anyone with a personal computer and a modem can search 1,000 newspapers for references to anything that comes to mind, from kiwifruit to quantum physics. Will it soon be possible for every person to carry a copy of the Library of Congress, to model the weather, to weigh alternative business strategies or to checkmate Gary Kasparov?

FURTHER READING

FIELD-EFFECT TRANSISTORS IN INTEGRATED CIRCUITS. J. T. Wallmark and L. G. Carlstedt. John Wiley & Sons, 1974.

THE PHYSICS OF VLSI SYSTEMS. R. W. Keyes. Addison-Wesley Publishing, 1987.

CAN WE SWITCH BY CONTROL OF QUANTUM MECHANICAL TRANSMISSION? Rolf Landauer in *Physics Today*, Vol. 42, No. 10, pages 119–121; October 1989.

LIMITATIONS, INNOVATIONS, AND CHALLENGES OF CIRCUITS AND DEVICES INTO A HALF MICROMETER AND BEYOND. M. Nagata in *IEEE Journal of Solid-State Circuits*, Vol. 27, No. 4, pages 465–472; April 1992.

THE FUTURE OF SOLID-STATE ELECTRONICS. R. W. Keyes in *Physics Today*, Vol. 45, No. 8, pages 42–48; August 1992.

Single Electronics

Is electric current the motion of individual electrons or the continuous flow of a fluid of charge? Recent experiments confirm both ideas and may lead to novel electronic devices

by Konstantin K. Likharev and Tord Claeson

What is the smallest amount of electric charge that can sit on the head of a pin? The answer may surprise. A pin, like almost everything else, consists of electrons, protons and neutrons. Each proton holds one fundamental unit of charge (denoted as +e), and each electron carries a similar negative charge (–e); the neutron has no charge. It may seem that the total charge on the head of the pin could be found by counting the number of protons, subtracting the number of electrons and ignoring the neutrons. Using this logic, the smallest nonvanishing amount of charge should be +e or –e. But recent experiments demonstrate otherwise. They clearly show that the charge on the head of a pin can be equal to a fraction of the charge of an electron, for example, +0.5e or –0.1e.

Whereas counting the number of angels on a pin has yet to yield anything of practical value, measuring charge in small structures has provided fascinating tests of physics as well as several potential applications for electronics. The tests were performed with extremely small devices, typically about 30 billionths of a meter in width, which is comparable to the diameter of a virus. By studying the behavior of such minuscule components at temperatures of a fraction of a degree above absolute zero, researchers have developed a better understanding of how electric charge moves through materials. Whereas charge passes through thin insulators in discrete amounts (like water dripping from a faucet), it travels through ordinary conducting materials in a continuous way (like water flowing in a pipe).

KONSTANTIN K. LIKHAREV and TORD CLAESON have collaborated for several years to develop single electronics. Likharev, who is now professor of physics at the State University of New York at Stony Brook, spent most of his career at Moscow State University. He first entered the institution as an undergraduate, then earned his Ph.D. there and stayed on as a research scientist. For three years, Likharev directed the Laboratory for Cryoelectronics at the university. Since 1982 Claeson has been professor of physics at Chalmers University of Technology in Göteborg, Sweden. He obtained his undergraduate degree and Ph.D. from Chalmers, and he has helped to establish a Swedish facility for fabricating nanometer-scale solid-state devices.

These discoveries have led to the invention of various novel devices that control the movement of individual electrons in solids. The single-electronic devices may serve as transistors and other computational elements. Although the technology is still in its infancy, we believe it will play a major role in the future of electronics.

The field of single electronics emerged from investigations of a device known as a tunnel junction [*see illustration on page 98*]. It consists of two electrodes—pieces of a conducting material—separated by a thin layer of an insulating material, as thin as about one nanometer (a billionth of a meter). Although electrons move freely in conductors, they do not readily enter insulators. Yet according to the laws of quantum mechanics, electrons have a small chance to pass, or tunnel, through the thin insulating layer.

If a voltage is applied across the junction, electrons will prefer to tunnel in one particular direction through the insulator. Hence, they will carry some electric current through the junction. The magnitude of the current depends on both the thickness of the insulating layer and the material properties of the conducting electrodes.

By the middle of the 1980s physicists understood many properties of tunnel junctions. A few physicists suspected, however, that tunnel junctions would exhibit interesting behavior if they were made smaller than 100 nanometers or so.

In early 1985 Dmitri Averin and one of us (Likharev) tried to predict the behavior of very small tunnel junctions with superconducting electrodes. (Superconductivity is a property of some materials cooled to temperatures close to absolute zero. Notably, superconductors do not resist the flow of electric current.) Averin and Likharev hoped to tackle the problem by applying ideas of quantum theory. They derived some equations that described the device, but the equations were so complicated that they could not have been solved without a struggle.

It was decided, therefore, to examine the easier case of a small tunnel junction with electrodes made from ordinary conductors. For this problem, the equations could be readily solved and gave a completely unexpected result. They showed that if a constant electric current is passed through a junction, it will induce a voltage that oscillates periodically in time. Most remarkably, these periodic oscillations would have a frequency simply equal to the current divided by the charge of an electron. The frequency is totally independent of any other parameters of the system. A more careful examination of the equations has shown that each oscillation represents the response of the device as a single electron tunnels through the insulating layer. The phenomenon is now known as single-electron tunneling (SET) oscillations.

To understand this effect, one must appreciate how electric charge moves through a conductor, say, an ordinary aluminum wire. An electric current can flow through the conductor because some electrons are free to move through the lattice of atomic nuclei. Despite the motion of the electrons, any given volume of the conductor has virtually no net charge because the negative charge of the moving electrons is always balanced by the positive charge of the atomic nuclei in each small fraction of the conductor. Hence, the important quantity is not the charge in any given

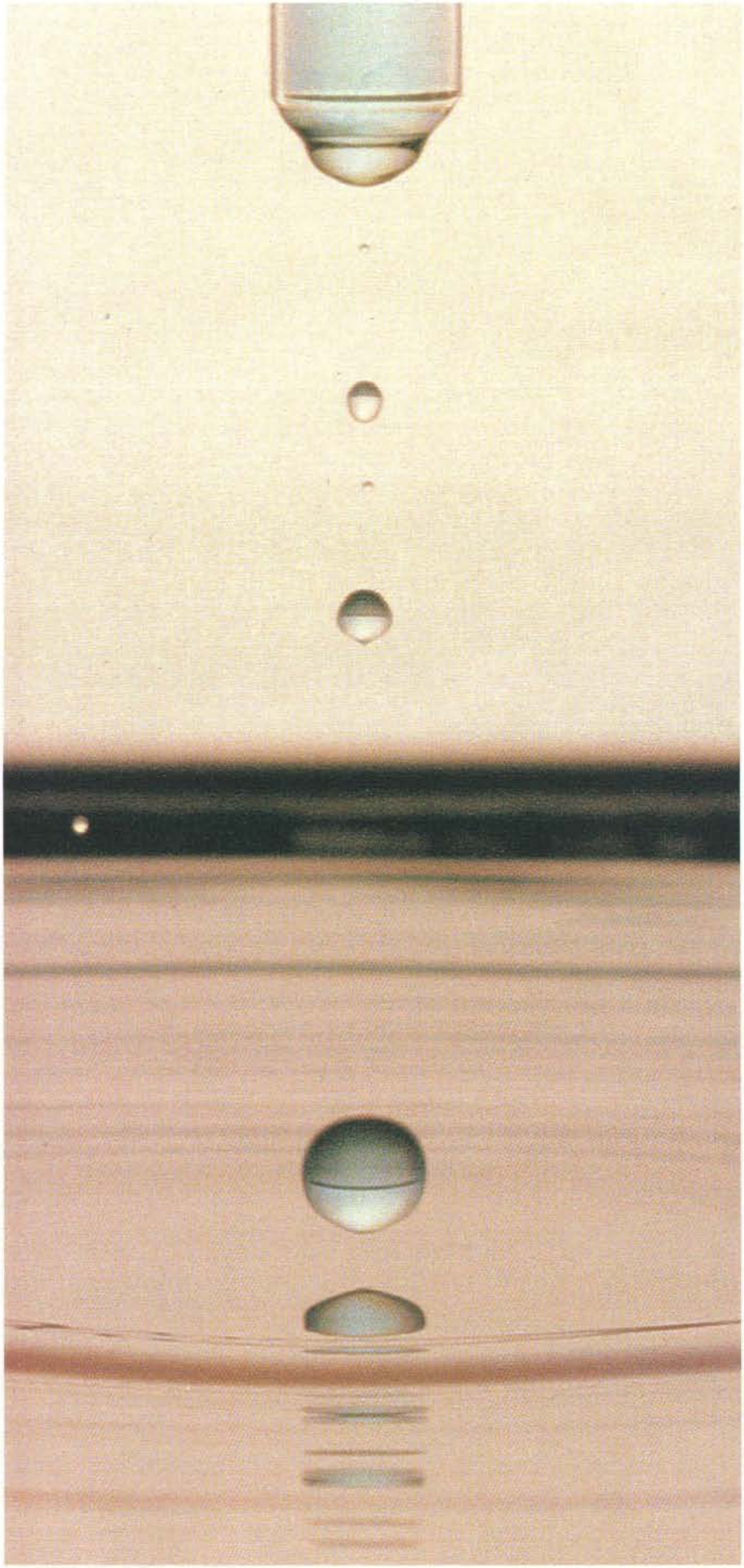
Reprinted from the June 1992 issue

volume but rather how much charge has been carried through the wire. This quantity is designated as the "transferred" charge. What is most surprising, this charge can have practically any value, even a fraction of the charge of a single electron.

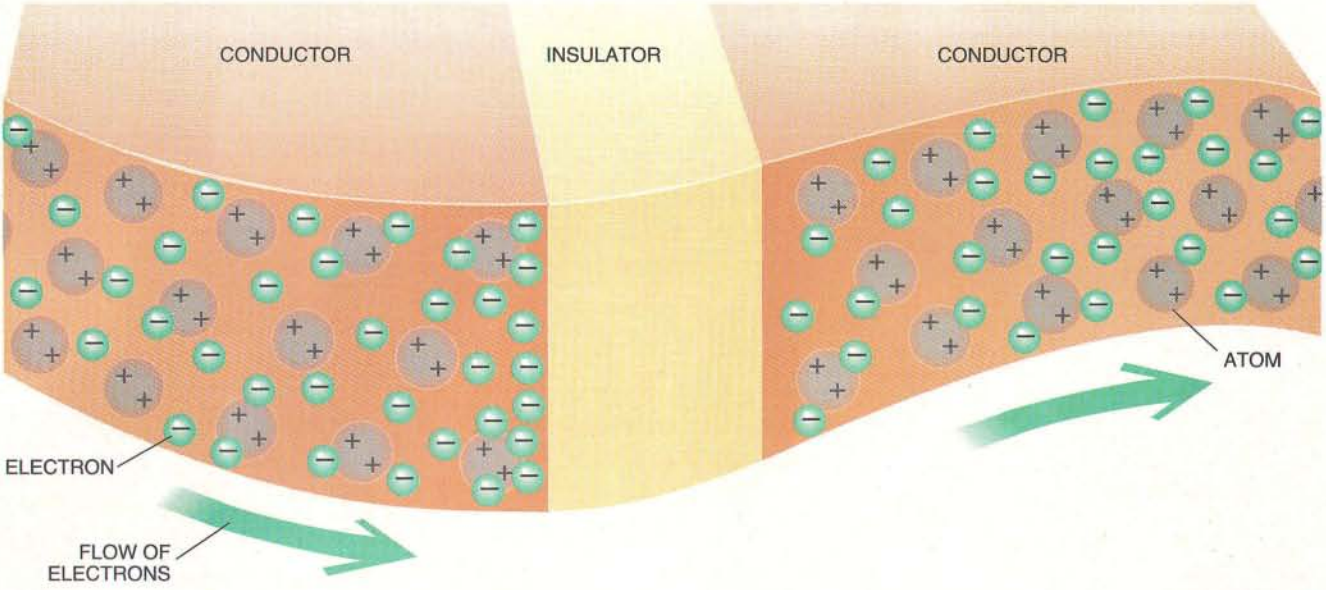
Although this concept may seem counterintuitive at first, it can be explained in simple terms. The transferred charge actually has little to do with counting single electrons or protons. This charge is proportional to the sum of shifts of all the electrons with respect to the lattice of atoms. Because the electrons in a conductor can be shifted as little or as much as desired, this sum can be changed continuously, and therefore so can the transferred charge.

If an ordinary conductor is interrupted by a tunnel junction, electric charge will move through the system by both a continuous and a discrete process. As the transferred charge flows continuously through the conductor, it will accumulate on the surface of the electrode against the insulating layer of the junction (the adjacent electrode will have equal but opposite surface charge). One can imagine this surface charge Q as a slight continuous shift of the electrons near the surface from their equilibrium positions. On the other hand, quantum mechanics shows that the tunneling can only change Q in a discrete way: when an electron tunnels through the insulating layer, the surface charge Q will change exactly by either +e or –e, depending on the direction of tunneling. The interplay between continuous charge flow in conductors and discrete transfer of charge through tunnel junctions leads to several interesting effects.

These phenomena can be observed when the tunnel junctions are very small and the ambient temperatures are very low. (Low temperatures reduce thermal fluctuations that disturb the motion of electrons.) In this case, if the charge Q at the junction is greater than +e/2, an electron can tunnel through the junction in a particular direction, subtracting e from Q. The electron does so because this process reduces the electrostatic energy of the system. (The energy increases in proportion to the square of the charge and does not depend on the sign of the charge.) Likewise, if Q is



WATER DROPLET forms at the end of a glass pipe and falls—a process analogous to the way charge moves through single-electronic devices.



TUNNEL JUNCTION consists of two conducting electrodes separated by a thin layer of insulating material. When a voltage is applied to the device, electric charge creeps through the insulator, creating a weak current.

less than $-e/2$, an electron can tunnel through the junction in the opposite direction, adding e to Q, and thus again decrease the energy. But if Q is less than $+e/2$ and greater than $-e/2$, tunneling in any direction would increase the energy of the system. Thus, if the initial charge is within this range, tunneling will not occur. This suppression of tunneling is known today as the Coulomb blockade.

Curiously, physicists first noticed and studied Coulomb blockade more than 40 years ago. In the 1950s and 1960s Cornellis Gorter of the Kamerlingh Onnes Laboratory, Hans-Rudi Zeller and Ivar Giaver of the General Electric Research Center and John Lambe and Robert Jaklevich of the Ford Motor Company observed this effect in thin metallic films and explained it. Meanwhile Igor Kulik and Robert Shekhter of the Kharkov Institute of Low Temperature Physics devised its comprehensive theory for one particular system. It seems, however, that until the mid-1980s no one appreciated the concept of continuous charge transfer in metals, and no one was aware of the simple condition for the Coulomb blockade. Once these ideas were understood, it was straightforward to discover new phenomena in small tunnel junctions.

For example, what happens if the junction is connected to a source of constant current? If the surface charge Q is zero initially, then the system is within the Coulomb blockade limits, and tunneling is suppressed. Therefore, the current flowing from the source through wires will start to change the charge Q continuously. For convenience, assume that the deposited charge rate is positive rather than negative. If the charge reaches and slightly exceeds $+e/2$, tunneling becomes possible. One electron will then cross the junction, making its charge slightly greater than $-e/2$. Hence, the system is within the Coulomb blockade range again, and tunneling is not possible. The current continues to add positive charge to the junction at a constant rate, and Q grows until it exceeds $+e/2$ again. The repetition of this process produces the single-electron tunneling (SET) oscillations: the voltage changes periodically with a frequency equal to the current divided by the fundamental unit of charge, e.

In this respect, charge flows through the tunnel junction like water leaking from a faucet. Initially, the charge collects in the junction just as water forms a droplet on the faucet. And indeed, when enough charge builds up, some of it pushes through the junction just as when the droplet reaches a certain size, it will fall. Yet whereas the water droplets can vary in size, the amount of tunneling charge is quantized; it always equals e.

With a couple of plausible assumptions, the formula for the frequency of SET oscillations can be derived from classical physics, because it describes behavior of electrons as particles. Nevertheless, a deep connection exists between SET oscillations and one type of oscillations predicted by quantum theory. In many situations, the electron behaves like a wave rather than a particle. The electron wave has a characteristic frequency, which is equal to the energy of the electron divided by Planck's constant. In 1962 Brian Josephson (then a student at the University of Cambridge) showed that in large tunnel junctions with superconducting electrodes, this relation can be observed experimentally as a fundamental relation between applied voltage and the frequency of oscillations in the current flowing through the junction. In the mid-1980s James Lukens and his co-workers at the State University of New York at Stony Brook used these "Josephson junctions" to demonstrate that the Planck-Bohr-Josephson relation is accurate to at least 16 decimal places. We believe the classical equation for SET oscillations will prove to be equally exact.

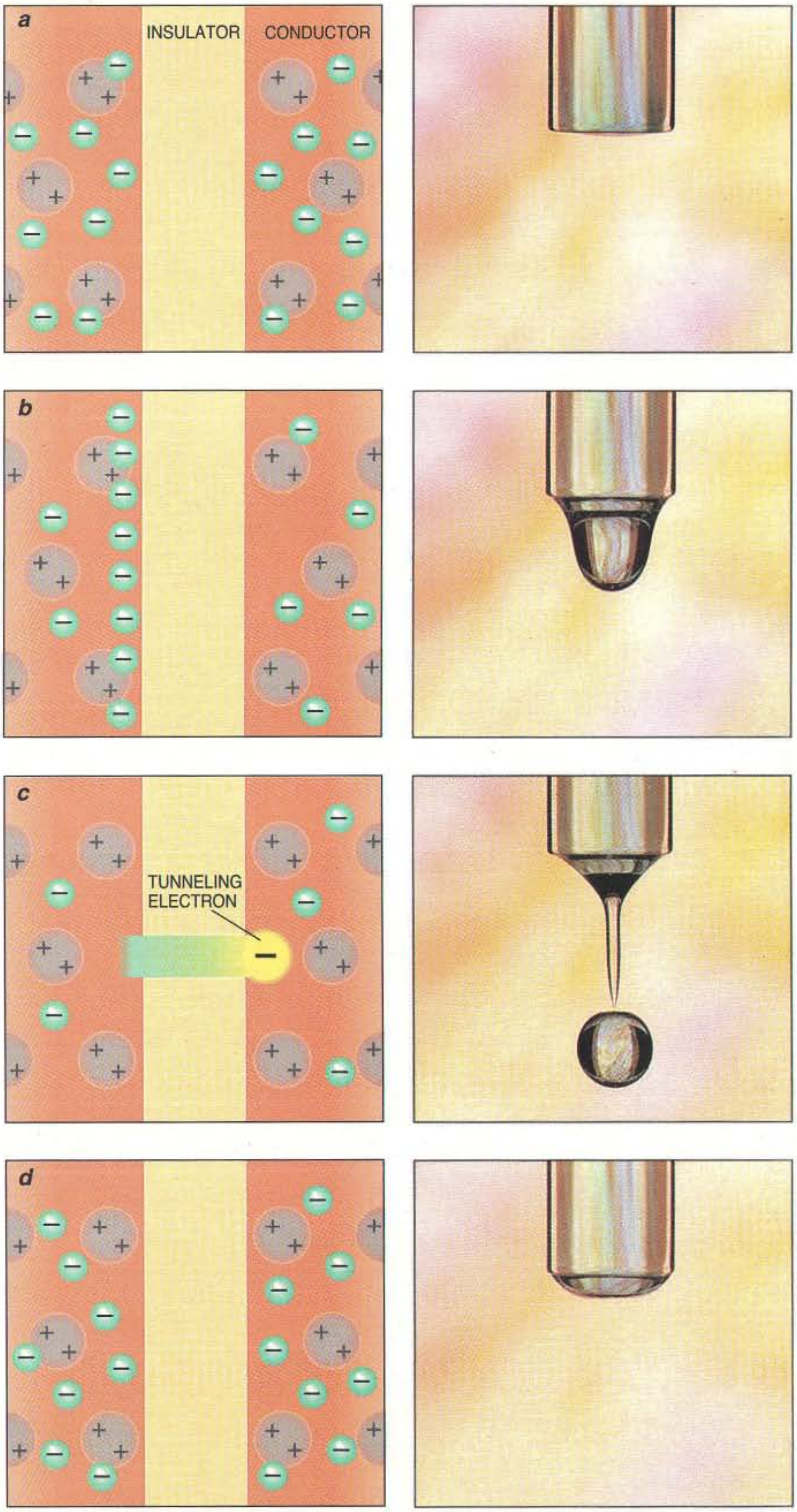
To produce SET oscillations, physicists must fabricate tunnel junctions of a very small area and cool them enough to ensure that the thermal energy does not influence tunneling. For example, the device must be cooled to temperatures of about a tenth of a degree above absolute zero if the junction is 100 nanometers in length and width.

Low-temperature refrigeration and small-scale fabrication are not the only experimental problems that must be overcome. To apply current and voltage to a tunnel junction, one must attach metallic wires to it. Unfortunately, the wires pick up quantum fluctuations of the electromagnetic fields that exist everywhere. These fluctuations can jam the single-electron effects completely. One of the simplest ways to eliminate the jamming is to connect several tunnel junctions in series (end to end). In such an arrangement the junctions defend one another from the fluctuations. This idea was implemented in 1989 by a team of Swedish and Soviet scientists. The team—which included Per Delsing of Chalmers University of Technology in Sweden, Leonid Kuzmin of Moscow State University and the authors—made the junctions by modifying the so-called suspended mask technique for nanofabrication. Several investigators contributed to the invention of the technique, which was perfected by Gerald Dolan of AT&T Bell Laboratories.

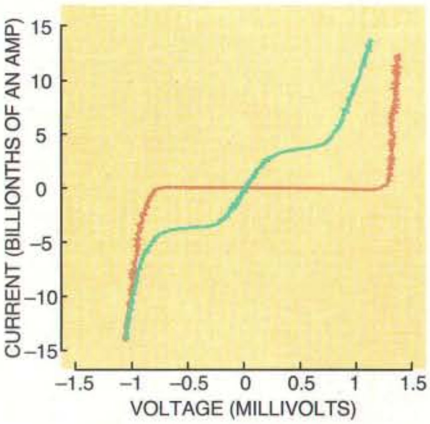
To start, an insulating substrate should be coated with two layers of special organic materials, both thinner than one micron. We used a copolymer of polymethylmethacrylate (PMMA) and polymethacrylic acid (PMAA) for the bottom layer, whereas the top layer was pure PMMA. The coated substrate is then placed into the vacuum chamber of a scanning electron microscope. Although the instrument is mainly designed for imaging applications, it is useful for device fabrication because it produces a beam of electrons a few nanometers in diameter, and the beam has enough energy to break polymer bonds. A computer directs the beam to sweep over the polymer coating in a desired pattern. In our experiments the pattern was a chain of closely located rectangular windows, 200 by 80 nanometers. The sample is then developed in toluene, which removes PMMA from the exposed regions and the underlying copolymer from a somewhat larger area. This process leaves a PMMA layer with open windows that is suspended above the substrate.

At this stage, the sample is placed into another vacuum chamber and is subjected to a beam of aluminum atoms. This beam deposits a thin aluminum film on the PMMA layer and on the parts of the substrate under the open windows. (These aluminum islands will ultimately become the first conducting layer of the tunnel junction array.) Then a small amount of dry oxygen is allowed into the chamber, forming a very thin layer of aluminum oxide on the surface. (The aluminum oxide will become the insulating layer of the junction.) The oxygen is pumped out again, and a second layer of aluminum is deposited from another angle. In this way, the islands of the first and the second layer overlap, forming an array of tunnel junctions, each close to 80 by 60 nanometers.

As our team began fabricating arrays of tunnel junctions, we also considered how we might best measure SET oscillations. The expected power of the oscillations would be very low. To make



SINGLE-ELECTRON tunneling through a small junction is similar to a water droplet falling from a glass pipe. Initially, the interface between the electrode and the insulator has no charge (*a*). As electrons flow into the electrode, charge builds up on the surface of the conductor (*b*). When enough charge collects there, an electron will pass through the insulator and thereby diminish the surface charge (*c*). The process will repeat (*d*) if the charge is replenished.



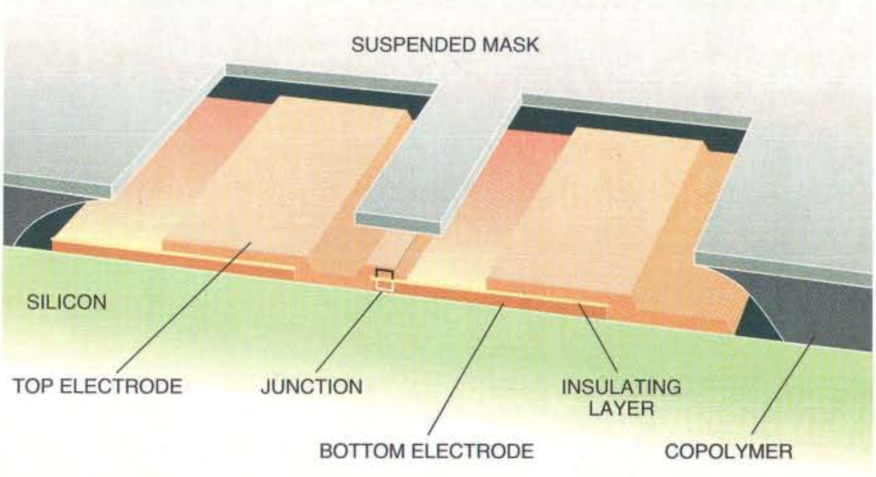
CURRENT through an array of 25 tunnel junctions is nearly zero (*orange line*) between −0.8 and +1.2 millivolts because of the Coulomb blockade. Yet when the array is exposed to microwaves, the blockade is suppressed, and the current increases in steps (*blue line*). The height of the steps equals the charge of an electron times the frequency of the microwaves.

direct measurements, we would need extremely sensitive high-frequency detectors. A simpler way to detect the oscillations is to irradiate the sample with microwaves. The radiation interacts ("mixes") with the SET oscillations in a way that can be easily observed. Specifically, as the current through the junctions increases from zero, the voltage should increase until the current equals e times the frequency of the microwave radiation. As the voltage continues to increase, the current remains constant for a while and then increases. As a result, a plot of current versus voltage should show horizontal steps at a current equal to the frequency times +e (or −e). In July 1989 we observed the expected steps [*see illustration above*].

Our results were soon confirmed by a Dutch-French collaboration of researchers working at Delft University of Technology and the Center of Nuclear Research at Saclay. These groups used two different types of junction arrays, in which the microwave field is applied to central electrodes rather than at the edges of the array. With these devices, the current-to-frequency relation was tested with an accuracy close to 0.1 percent. Moreover, in 1990 Bart Geerligs and Hans Mooij of the Delft group confirmed a theory by Averin and Arcadii Odintsov, which shows how the accuracy might be improved. We believe that in the near future the current-to-frequency ratio will be measured with an error below at least one part per million—and probably much better.

Phenomena similar to single-electron oscillations should also occur in tunnel junctions that have superconducting electrodes. In 1982 Allen Widom, Terry Clark and their collaborators at the University of Brighton first suggested a rudimentary theory for such phenomena in superconducting tunnel junctions. Then, in 1984, Alexander Zorin of Moscow State University and one of us (Likharev) developed a more refined and realistic theory.

In superconductors, each electron has a partner, forming what is known as a Cooper pair. Consequently, if a tunnel junction is made from two superconducting electrodes, the electrons tend to tunnel through the insulator in pairs. Theory predicts, therefore, that the voltage of a very small superconducting tunnel junction will oscillate in time with a frequency equal to the current divided by 2e (rather than e). Likharev and Zorin called this particular effect Bloch oscillations. In February 1991 Kuzmin, David Haviland and their colleagues at Chalmers found reliable evidence for Bloch oscillations in a small junction formed between two thin films made of a superconducting alloy.



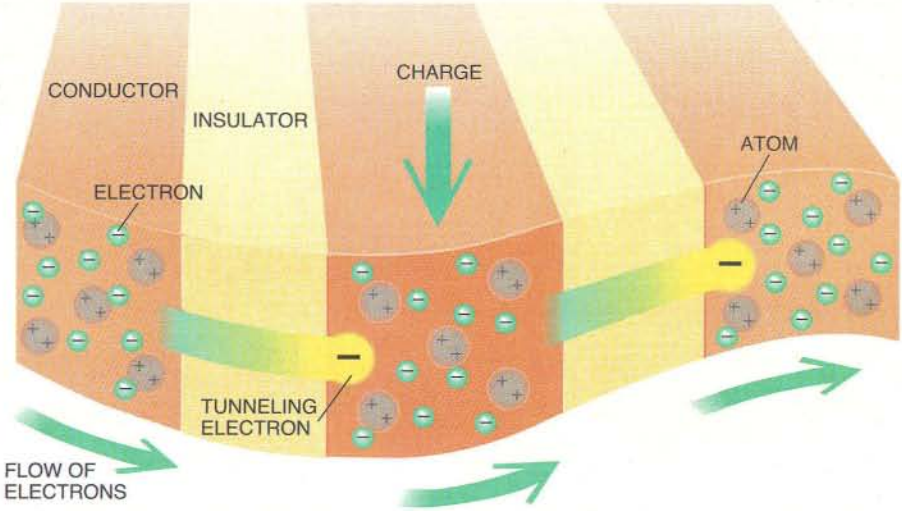
TINY TUNNEL JUNCTIONS were made using the suspended mask technique.

The SET and Bloch oscillations can be considered as the ordering ("correlation") of single-electron or Cooper-pair tunneling events in time. Another type of correlation of these events is also possible and is important for future applications. This type of correlation occurs, for example, in a device consisting of two tunnel junctions that share a middle electrode. If a constant voltage is applied across the device, the tunneling events in each of the junctions should be mutually correlated. Specifically, whenever an electron tunnels through one of the junctions, a second electron passes through the other junction nearly simultaneously.

Most important, as Averin and Likharev showed in 1985, this correlated tunneling can be controlled if electric charge is continuously injected into (or removed from) the electrode that connects the two junctions. If the charge injected into the electrode is close to zero or a multiple of e, the device maintains a considerable Coulomb blockade: no current flows through the junction if the voltage applied to the system is lower than a particular threshold. But if the charge of the middle electrode is either +0.5e or −0.5e (or even +1.5e, +2.5e and so on), the Coulomb blockade is completely suppressed: electrons tunnel through the system, and a current can be induced by even a small applied voltage. Thus, the current through the device can be controlled by changing the charge of the middle electrode just like the current through a usual semiconductor transistor, but this control can be achieved by much weaker charge signals. For this reason, the two-tunnel-junction device is now known as the single-electron transistor.

Fortunately, the correlated tunneling effect is not sensitive to quantum fluctuations of the ambient electromagnetic field, and so the device can be connected easily to a voltage source and measuring instruments. For this reason, Kuzmin and Likharev were able to conduct experiments on the single-electron transistor very rapidly after the device was conceived. On March 6, 1987, they reported their results to a Soviet journal. (They later learned that, on the same day, Theodore Fulton and Gerald Dolan of Bell Labs had submitted a paper on almost similar observations to an American journal.) These two works marked the beginning of experimental single electronics.

During the past five years, physicists have devised many types of single-electron transistors and developed more



SINGLE-ELECTRON TRANSISTOR can switch on or off the flow of billions of electrons per second when the charge on the middle electrode is changed by only half the charge of an electron.

complex single-electronic circuits. The behavior of all these circuits can be fully explained using what is now called the Orthodox theory, which the Moscow group formulated in 1985. Hence, the Orthodox theory can be used to analyze possible applications of single electronics.

In the near future, we foresee at least three important applications. First, the technology should provide a new standard for measuring small currents. We expect an accuracy of better than one part per billion, about 1,000 times better than existing systems. Second, the technology promises supersensitive electrometers—instruments that would measure charges as small as one ten-thousandth of e, which is almost a million times better than the resolution of commercially available instruments. (Such electrometers have already been used by researchers at Saclay and at Bell Labs to measure single-electron tunneling effects in other devices.)

But the most important prospect for single electronics is its application to digital integrated circuits. Currently computer chips can have a density of some 10 million devices per square centimeter. Although this achievement is impressive, computer designers continue to demand chips with ever higher device densities. But the conventional electronic devices are approaching their limit. Most agree that any further reductions in size will require radically new ideas, and single electronics is ranked highly on this issue.

In circuits based on single electronics, bits of information can be represented as the presence or absence of individual electrons. This technology may make it possible to pack up to 10 billion electronic devices on a one-square-centimeter chip (for that, the dimensions of tunnel junctions could be reduced to about 10 nanometers). Such a circuit would operate at four kelvins, a temperature that is acceptable for some applications. Still, the development of such circuits will require a considerable investment of time and effort.

Conceivably, single-electron circuits could be made even more dense, but such attempts will encounter two problems. First, researchers need to pioneer techniques to fabricate complex structures whose smallest dimension is less than 10 nanometers. Second, they must determine how single-electron effects change as the dimensions of the device decrease to the atomic scale.

To solve this second problem, Averin and Alexander Korotkov of the Moscow State University revised the Orthodox theory in 1989. They recognized that if a device is small enough, the laws of quantum mechanics play a larger role in the system. In such cases, the energies of the electrons are quantized: that is, they can only equal certain discrete values. Yet they showed that the energy quantization does not exclude correlated tunneling, and indeed the effects can coexist peacefully.

To verify this conclusion, several groups have performed experiments in which electrons are confined within tiny structures made of semiconducting material. Such structures, which are known as quantum dots and quantum wells, can hold a "puddle" of a few hundred free electrons. Both devices allow electrons to tunnel in or out through well-defined energy barriers. In both types of experiments, physicists have observed single-electron effects and energy quantization together, confirming that the phenomena can coexist. Paul McEuen and his colleagues at the Massachusetts Institute of Technology and the IBM Thomas J. Watson Research Center were the first to demonstrate this coexistence in quantum dots; Bo Su and his co-workers at Stony Brook and at Bell Labs have obtained similar results using quantum wells.

Several fundamental issues are still to be solved. For one, investigators have not yet tackled the issue of how energy quantization influences single-electron charging effects in circuits that consist of many quantum dots or quantum wells. A preliminary analysis shows that in such structures an electron can behave simultaneously as a wave and as a particle. If this prediction turns out to be true, it will be very important for fundamental physics.

Another unresolved issue is whether the correlated transfer of electrons requires tunneling. Averin and Likharev have recently given arguments that it does not. Such correlated flows may be achieved in narrow channels, where the electrons can propagate only in one direction. Forthcoming experiments should reveal whether such channels can be fabricated.

Regardless of these new issues, single electronics has already advanced our understanding of how electrons behave in materials. We believe that a contribution has been made as well to what could be called the psychology of physics. The discovery of correlated tunneling has given us an example of a conceptually simple and fundamental effect that remained undiscovered until the mid-1980s. It has shown that Newton's "ocean of undiscovered truth" is still full and is waiting for scientists armed with not much more than a bit of imagination and creativity. This is something to be remembered by adherents of theoretical supersophistication and advocates of multibillion-dollar experiments.

FURTHER READING

CORRELATED DISCRETE TRANSFER OF SINGLE ELECTRONS IN ULTRASMALL TUNNEL JUNCTIONS. K. K. Likharev in *IBM Journal of Research and Development,* Vol. 32, pages 144–158; January 1988.

SINGLE ELECTRONICS. K. K. Likharev in *Granular Nanoelectronics.* NATO ASI Series B, Physics Vol. 251. Edited by David K. Ferry et al. Plenum Press, 1991.

SINGLE CHARGE TUNNELING. Edited by Hermann Grabert and Michel H. Devoret. Plenum Press, 1992.

Products and Services for Computer Networks

Tired of the constraints of space and time? Intelligently designed network products that understand the needs of individuals will set us free

by Nicholas P. Negroponte

During a dinner party, you might wink across the table at a close friend. By applying a slightly perverted measure of bandwidth, we can claim that your wink delivers a mere one bit. Yet that bit carries an enormous amount of information; if asked, you might need more than 100,000 bits to explain the content of the message to a stranger. In a real sense, you achieved data compression in excess of 100,000 to one. With that degree of compression, we could send high-definition television (HDTV) over telephone lines at the sluggish rate of only 1,200 baud (bits per second).

The success of the wink cannot be understood through the normal study of network efficiency. When sizing up networks, we tend to look at their channel capacity, without due consideration of the computational abilities of the transceivers. The wink works for a contrary reason: the transmitter and the receiver share a body of common knowledge and experiences, and they have the intelligence to put the act of winking into context. At the moment, we are devoting almost no effort to building networks with similar cunning.

If we are looking for creative applications of future networks, then we cannot isolate thoughts about channel mode and capacity—flat estimates of how much information can travel how fast—from the computational resources at the ends of the network. The propositions supporting that truth seem counterintuitive to some people; indeed, many of my colleagues do not yet accept them. Nevertheless, to prove my point, I will look at the relevance of this truth to three domains of products and services: entertainment, transactions and electronic personal surrogates.

NICHOLAS P. NEGROPONTE is the founder and director of the Media Laboratory of the Massachusetts Institute of Technology, where he is also professor of media technology. He has written and lectured extensively about the promise of computer technology, human-machine interfaces and the changing nature of communication. Negroponte studied architecture at M.I.T., receiving his master's degree in 1966. Within a year he had founded the Architecture Machine Group at the same institution. He has been a visiting professor at the University of Michigan, the University of California at Berkeley and Yale University.

First, however, let us debunk bandwidth. Many networking professionals argue that networks will be useless without the broad bandwidths made possible by optical fiber lines. They say that we must push to replace copper wires with fiber; only then can we have broadband networks capable of sending data at sufficiently high speeds.

Yet, ironically, broadband systems are inevitable. The cost and maintenance for fiber lines will be so much less than for copper ones that fiber will be installed even without the need to accommodate wideband services. The rush to justify gigabit services is merely a historical anomaly; today is a time of crossover between the falling cost of fiber and the generally steady cost of copper. The installed base of fiber will grow no matter what new services we offer. The problem is that its natural growth is too slow for those who can imagine the benefits of bandwidth—or who might profit from them.

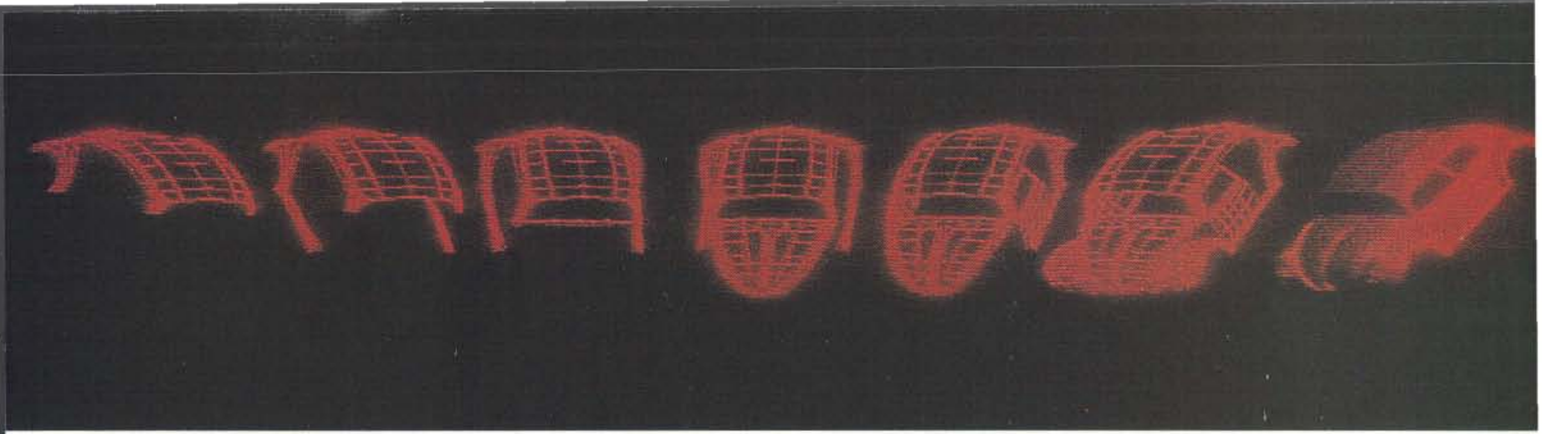
By trying needlessly to justify the inevitable, many network products and services being proposed now are contrived or, worse, made with total disregard for concurrent advances in computing. People with a vested interest in the network are saying that we need big data pipes connecting our machines because we will be naive about how to process and store information. One such claim, for example, is that we will need broadband networks to transmit video. In fact, we already know that video can be delivered over the T1 (1.5-megabit-per-second) twisted-pair copper wires existing today. The real products and services of the future will come from imaginative applications of both channel and computing capacity, not from either alone.

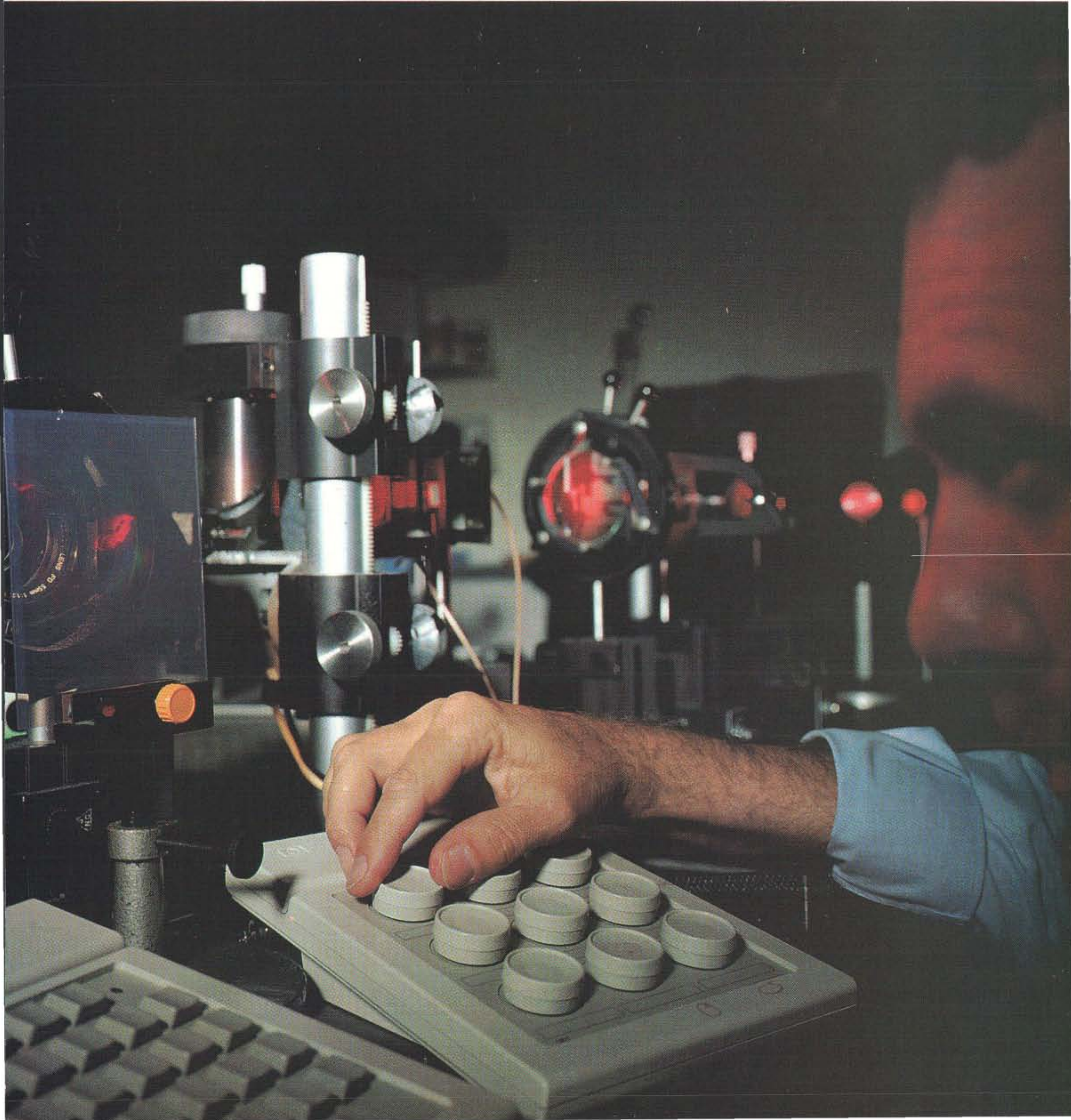
Postmodern man is potentially a nomad. An educated guess is that the authors in this issue of *Scientific American* spend on average more than half of their time traveling. According to one theory, if we had good teleconferencing and extraordinarily easy to use broadband services, most travel would be unnecessary—we could all stay home, preserve the ozone layer from jet exhausts and avoid the wear and tear on our bodies. Yet staying at home would not be the most interesting consequence of better networks.

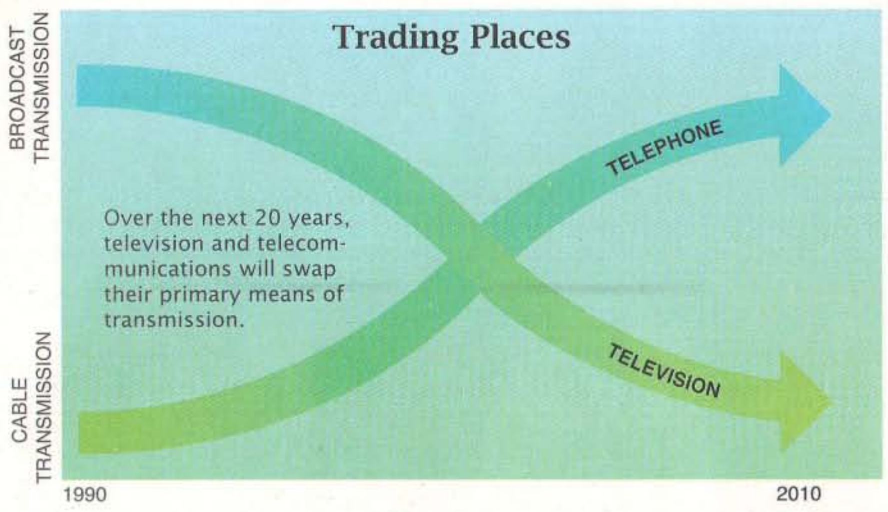
The often overlooked point is that we can freely move about precisely because we have the electronic means to stay in touch with our home base. No one needs to be more than a telephone call, data transfer or fax transmission away from the office—indeed, from any office. What we network planners should be looking at is not the bits per second required to lower our frequent-

MOVING HOLOGRAMS created by computers in real time have been developed by the Media Laboratory of the Massachusetts Institute of Technology with support from the Defense Advanced Research Projects Agency and Rome Laboratory. As the technology for producing these images improves and is integrated into the information network, holograms may become important vehicles for presenting news and entertainment.

Reprinted from the September 1991 issue







Over the next 20 years, television and telecommunications will swap their primary means of transmission.

flyer mileage but the computational and telecommunications resources that will allow us to be more or less independent of space and time.

Independence of space and time is the single most valuable service and product we can provide humankind. This outcome need not translate itself into a nomadic existence, although the concept is not without interest. It can also allow us to remain in our pajamas as long as we please.

Travel embargoes during the Persian Gulf War resuscitated interest in teleconferencing, which undoubtedly will continue to gain appeal (if only by corporate fiat in response to gratuitous business traveling). The more fascinating developments, however, will come from new services that free you to wander and that create an electronic surrogate for you on the network with which others can communicate.

Those developments will of course hinge on storing and transmitting information, either through electromagnetic broadcasts or over optical fiber. Broadcast spectrum is scarce, whereas fiber, like computing power, is something we can just keep making more of. Those facts mean that the channels for distributing different types of information, as we know them today, will trade places. Most information we receive through the ether today—television, for example—will come through the ground by cable tomorrow. Conversely, most of what we now receive through the ground—such as telephone service—will come through the airwaves.

Two rules of thumb define how information should be distributed. First, use the broadcast spectrum to communicate with things that move: cars, boats, airplanes, wrist telecommunication terminals, dog collars, scuba tanks and the like. Second, deliver information to the desktop or living room by fiber.

It is already easy to see these rules in action. The popularity of cellular telephones and cable television initiated the trend, and it will not stop. Within 20 years, it will be perverse, if not illegal, to use satellites for broadcast television. The Federal Communications Commission (FCC) should put the television networks and independent stations on notice that their spectrum will slowly be withdrawn and encourage the cable operators and telephone companies to begin collaborating immediately.

Do you recall O. Henry's Christmas short story, "The Gift of the Magi," about an impoverished husband and wife? He wanted to buy a comb for her long, beautiful hair, and she wanted to buy him a gold chain for his pocket watch. She secretly cut and sold her hair to buy the chain, and he pawned his watch to purchase the comb. Something akin to this is happening with television and computers.

Television manufacturers are increasingly adding computing power to their receivers, and computer makers are putting more video into their workstations. Although this development may seem like a merger of the technologies, each side is disregarding the underlying utility of the other's techniques and aims. As a result, the television makers are cutting off their own access to the broader information delivery market, and computer manufacturers are pawning their license to be in the entertainment industry.

After considerable nationalistic fuss, the U.S. has emerged as the potential world leader in advanced television by finally opting to go digital. Europe and Japan are embarrassed by the development because they have chosen to take nondigital paths. Nevertheless, the U.S. has farther to go in its approach to HDTV because, at the moment, all the proposals to the FCC are sadly and uniquely focused on delivering better-resolution pictures. Digital television has much more to offer than higher picture resolution.

Two fundamental attitudes toward digital image transmissions in general and television in particular need to change before we will see creative thought in this area. First, we must think of images as scalable. Bigger pictures require proportionately more pixels to represent them and more information to describe them. Today both a four-inch video display and a 27-inch display use the same number of pixels, so the quality of the images changes with size.

Second, in the long run, model-based image transmission and encoding are better than transmission of pictures alone. Mathematical models of a scene can describe the spatial relations of the objects in it and maneuver them through space. The idea of capturing a picture with a camera is obsolete if one can instead capture a realistic model from which the receiver can generate any picture. For instance, from a real-time model of a baseball game, a fan watching at home could get the view from anywhere in the ballpark—including the perspective of the baseball.

The entertainment industry will be one of the first to capitalize on the new networking environment for products and services. In fact, some groundwork for expansion is already in place. Nintendo currently has an installed base in the U.S. of more than 30 million electronic game machines; they are found in more than 70 percent of all the homes with a child between the ages of eight and 12 years. In short, Nintendo is America's largest domestic computer presence and potentially the country's major force for educational change. (Seymour A. Papert of the Media Laboratory at the Massachusetts Institute of Technology is working with Nintendo to understand better what children already learn from electronic games.) On a network, these machines would change the face of our culture.

But Nintendo does not see itself as a computer maker; company executives maintain they are in the entertainment business. They are not suffering a lack of vision; everyone knows that entertainment is big business. Furthermore, it is likely to become very high tech and a major player in network products and services. Speaking more generally, as products and services continue to evolve, businesses will be more inclined to define themselves by the services they offer to customers and less by the technologies they use.

The first network entertainment service to become available is likely to be "movies on demand" because the technology for it will be developed the soonest. How often have you walked into a video store and rented a movie that you have already seen or that turned out to be something you would never want to watch? Movie buffs rent videos brilliantly, but the rest of us are unable to remember which Clint Eastwood film we have seen or to know from its title that *Silence of the Lambs* is a chilling movie about a psychopath. The in-store classifications of drama, action, comedy, foreign and so forth are of little help.

Imagine a different scheme in which you could look at a dozen playing card-size movie trailers, or previews, on your TV screen at the same time. Scalable video makes such a plan entirely possible. By modestly interacting with your TV to indicate your choices of actors or themes, and by relying on your TV's record of what you have already seen, you could easily select the perfect film from any of the 50,000 movies made in the Western world since the inception of cinema. Indeed, it would be only slightly more difficult to ask only for a movie that is similar to another one that you liked.

The network's task would be to store movies somewhere and to deliver them to you instantly. The storage task is not outrageous. For current VHS quality, one hour of movie is about five gigabits. Although the best commercial dynamic random-access memory (DRAM) chips today can hold only 64 megabits, capacities are growing exponentially. By 2000 it should be possible to store a film on five one-gigabyte chips. Just 250,000 chips will be able to store every Western movie ever made. Until then, all 50,000 movies could be stored on twice that many compact discs (CDs) or, in the still nearer term, on 50,000 videodiscs.

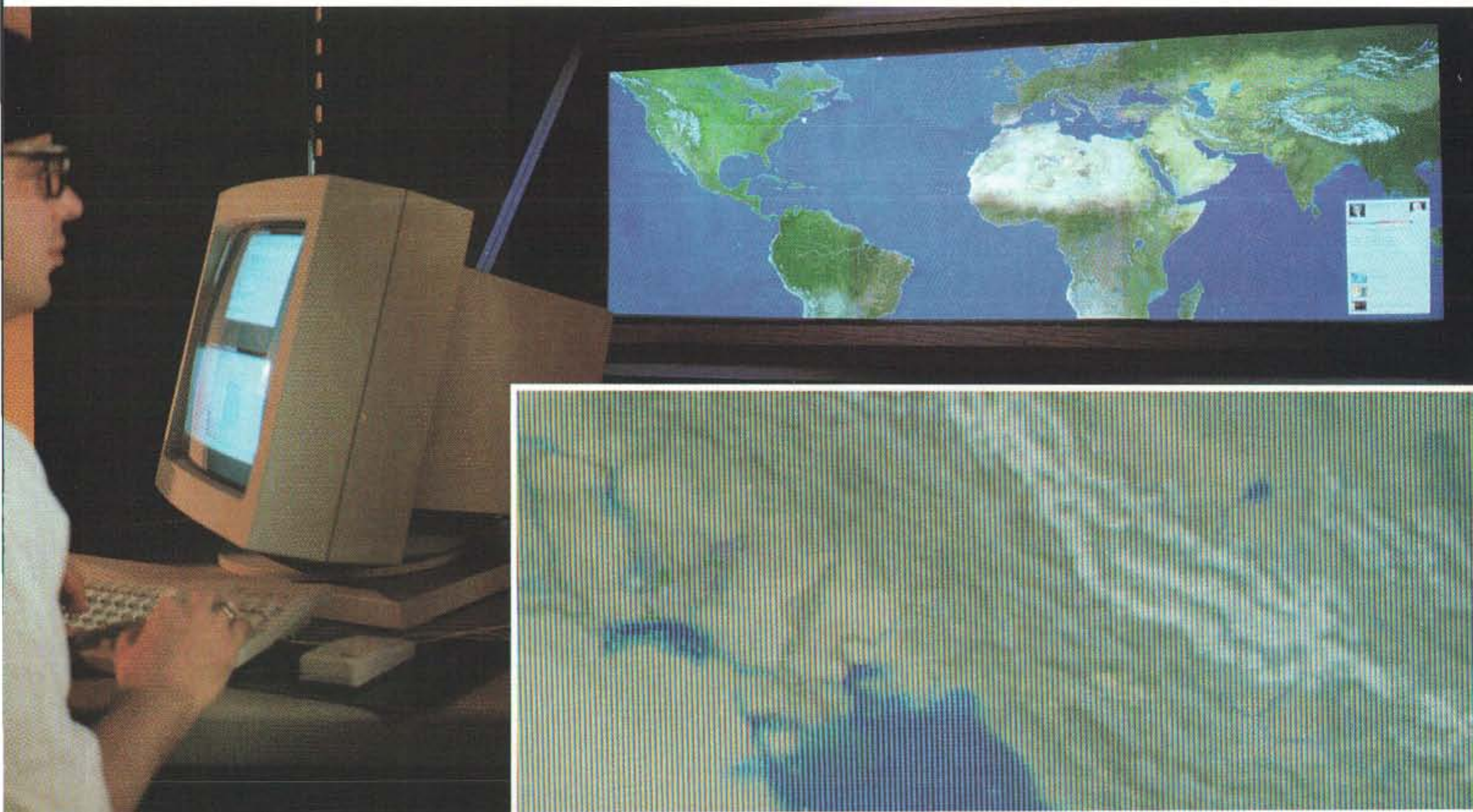
The movie-viewing customer is not the only one who would benefit from this scheme. When M.I.T.'s Media Lab was asked to investigate the technology for Columbia Pictures in 1983, the idea was to use the compact audio disc as a new form of packaged media that could be purchased for a price equal to that of rental. The idea appealed to movie companies because they could get a piece of the action for each copy sold to the public, instead of the one-time income from tapes sold to video rental stores. Today a standards organization, the Motion Pictures Expert Group, is helping to coordinate the efforts of IBM, Apple, Intel, Philips, Sony and other companies that have become interested in the development of compact-disc video.

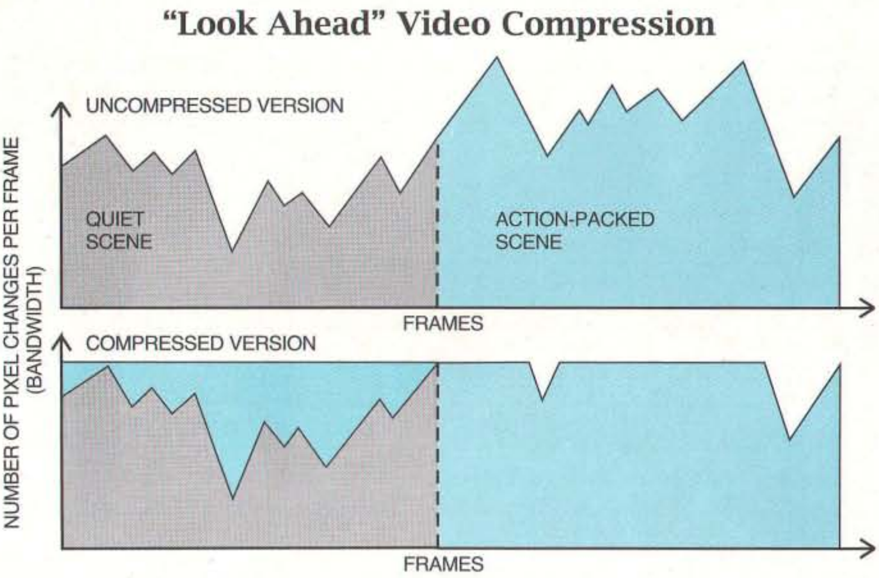
Because a movie would be encoded only once by its marketers and decoded millions of times by viewers, considerable effort and money can be applied to ensure the quality of the encoding process. Encoding need not be done in real time. Sophisticated data compression techniques can "look ahead" through a video and find scenes rich in information (that is, rapid rates of pixel changes); some of those "excess" data can be saved along with earlier scenes that have little information. By stuffing the "peaks" into the "valleys," one can maintain consistently high data compression. In this fashion, one hour of a movie not only fits on a CD but can be pushed through the CD's narrow-band audio port.

An even more interesting application of the same compression techniques is to bundle video data into efficient packets that can be transmitted rapidly. In this way, over a fiber running at a gigabit per second, one hour of video can be shipped in only five seconds. Therein lies a cogent example of how using both computer intelligence and bandwidth to deliver a service is superior to the traditional approach, which declares that video is a bandwidth hog requiring a continuous delivery stream.

In packeted form, video can be delivered to a memory-equipped receiver and viewed in an order different from

VERY HIGH DEFINITION TELEVISION being developed at the Media Lab offers almost photographic-quality images. The display has 2,000 vertical lines of resolution, compared with the 480 lines for standard television. Only in extreme close-up does a detail of the Persian Gulf (*inset*) begin to exhibit the poor quality seen on current television sets. The 6,000-pixel width of the display is achieved by tiling three 2,000-pixel monitors side by side.





"Look ahead" video compression is a technique that makes it possible to encode and transmit recorded video information efficiently. Prior to compression, the levels of information change greatly from frame to frame (*top*). This technique reduces the bandwidth required to deliver the signal by fitting the "peaks" into the "valleys" (*bottom*). After additional compression techniques are performed, one hour of video can be transmitted over a fiber in less than five seconds.

that in which it was sent. Related packets would not be sent sequentially but would be interlinked or indexed by content. Viewers could navigate up or down through this hierarchy of information. Someone watching a television news program, for instance, could push the "tell me more" button on the TV, and Dan Rather would elaborate on a particular story to various levels of depth—all the while providing substance that would otherwise have been relegated to the cutting-room floor.

Although television need not be a bit guzzler, holographic video is another story. Fifteen years ago in *Star Wars,* many of us saw R2D2 beam Princess Leia's three-dimensional image onto Obi-Wan Kenobi's floor. Cinematic special effects aside, real-time holography will not appear in living rooms for a few decades. The Media Lab, which is to my knowledge the only laboratory working on real-time holography, has achieved it for images measuring about one cubic inch. Those images are achieved with a massively parallel, multimillion-dollar computer, a roomful of electro-optic instruments, a spinning mirror and a heavy-duty laser. It is unclear how that technology will scale up to allow eight-inch-high football players to run around your house passing a half-inch football.

Look at the numbers. The Media Lab's current system uses a 50-megabit-per-second channel to generate a small image. Scaling up to a living room–size football game would require computing 45 terabits per second by brute force. Even if we subtract three orders of magnitude by using "smart" processing techniques, we are still looking at 45 gigabits per second.

The massive bandwidth consumption of holography suggests that it will be necessary for a receiver to get a computational model of an event rather than raw holographic data. The holographic images for the living-room football game could be generated from an accurate model. Information for manipulating that model would be sent over a fiber, but the image would actually be computed and created by the receiver—not captured by a camera.

Regardless of whether the model or the hologram is transmitted, substantial bandwidth is required—the bandwidth that only fiber can deliver. Although my football example is drawn from the world of entertainment, it is likely that the technology will first be developed in medical imaging for remote diagnosis in the near term and teleoperations in the longer term.

Information has value, but it is as perishable as fresh fruit. It varies from person to person, from moment to moment. And in some cases, the value can suddenly drop to zero: the knowledge that a seat on an airplane is empty becomes worthless when the plane pulls away from the jet way. But when information leads to a transaction, such as a booking for that empty seat, the value is less ephemeral. A service that facilitates transactions earns the gratitude of both the buyer and the seller. The service, in turn, is equally happy because it can probably bill both parties. Information transactions have so much appeal for both businesses and consumers that they are sure to be among the first network products and services.

Multimedia computing, which combines moving video, audio and text, is likely to be an important tool for transactions. It is anticipated to be a serious bandwidth consumer, but short-term multimedia solutions need not be. Customers at home could access some information at a remote site by narrowband communications. The rest could already reside in the home at the time of the transaction, having arrived there by regular mail or overnight delivery. (If you send three CDs by Federal Express Priority One service from Los Angeles to Boston, then the information transfer is greater than a modern 64-kilobit-per-second data line could provide during the same time.)

A simple example would be the use of a videodisc for summer-home rentals. As many as 54,000 houses could be pictured on the disc, which could be distributed to potential renters. Users would control the disc by specifying what they wanted in a summer house: the cost, the number of rooms, the seasonal availability, the location and views, and so on. The viewing system would display only the available houses that fit the criteria. In this case, videodisc technology created a product, but the network mixed the central and local computing services to make the transaction possible.

More generally, the problem and solution can be described as a kind of Yellow Pages. In this context, however, the term has several meanings that go beyond those you might associate with this venerable publication. For example, electronic Yellow Pages could be geographically coded so that a company or service could be related to who you are, where you are, where you are going and what might be along the way. Because this Yellow Pages would be an active part of the network rather than a static book on a shelf, it could contain up-to-the-minute information about its advertisers. It could include, for example, tonight's specials at your favorite restaurant or today's bargains at Bloomingdale's.

More than 95 percent of the services that advertise in the Yellow Pages use them as their only advertising. For that reason, the opportunity for adding layers of detail, daily information, photographs, video pitches and audio narratives will appeal to service providers. The customer would also welcome that information. The data base could be explored from totally original angles, such as "Who is serving fresh white truffles tonight?" Transactions that result from those queries are more like making a date than buying yogurt.

The opportunity is made-to-order for multimedia network products and services. Yet, it will require Federal Judge Harold Greene to relax some of his restrictions on the telephone companies that bar them from getting into the information-providing business. Simultaneously, the current information providers, particularly newspapers, will have to become more aggressive and imaginative in the electronic milieu.

Today print advertising falls into roughly three groups: persuasion by affiliation, convincing with facts and selling the specific. By affiliating their products with yachts, cowboys and good-looking people, cigarette and liquor ads are archetypes of the first category. Automobile and department store ads are an example of the second. The classified section of the newspaper embodies the third. In every case, the reader must more or less stumble on the relevant information: it does not percolate to the top of the information strata when the reader enters the market for a car or a summer suit.

The big change for print media is that advertising will become news and that persuasion by affiliation will die off. The network will be sufficiently personalized to anticipate your needs and interests and smart enough to find obscure auctions of fly-fishing equipment in the dead of winter. Advertisers will not necessarily be unhappy with such a network, because the broad reach is narrowed to a few who have a likely interest. Imagine how delighted General Motors and Nissan would be to have the opportunity to advertise to you specifically when you start looking for a new BMW.

Almost all services can distinguish themselves by becoming personalized. We are pleased to be recognized and catered to as individuals rather than being treated as a faceless part of the mob. Personal information systems, like the wink, are also extremely efficient. In this regard, the most desirable interaction with a network is one in which the network itself is invisible and unnoticeable. Planners often forget that people do not want to use systems at all—easy or not. What people want is to delegate a task and not worry about how it is done.

Calling someone on the telephone is an excellent example: the physical design of the handset, the brightness of the displays and the feel of the buttons should be utterly irrelevant. The caller simply wants to direct the telephone to contact so-and-so and to do something appropriate if that person is not available. Finding the person (or failing to do so) is only a part of the task; doing the appropriate thing is what matters and takes extraordinary personalization and intelligence.

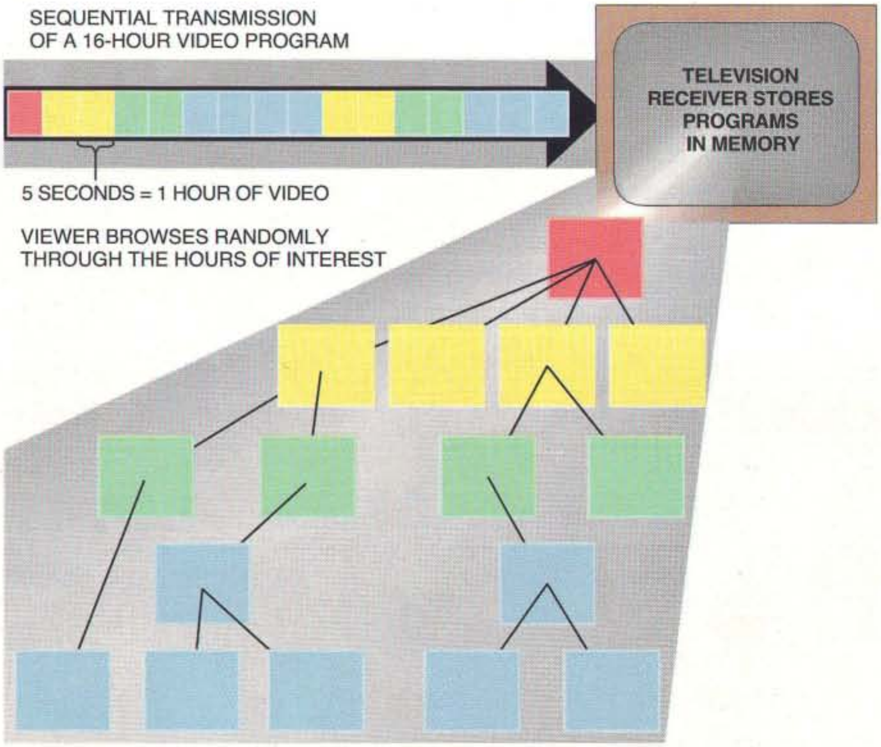
What constitutes knowing someone? Usually it is a shared body of knowledge, but not in the sense of two people knowing organic chemistry. Rather it is in the sense of two people sharing experiences. Each of us has a sphere of acquaintances, professional and personal, with whom we work and play. Similarly, we each have our own habits, patterns of life and idiosyncracies. Our jobs involve networks of people who assume different degrees of importance in our lives at different times. We know them because we are familiar with their ways, and they with ours.

What makes a system personal is its knowledge of such information. In a very real sense, the system becomes a bona fide member of your personal and professional community. Electronic personalization is by definition a mix of centralized and decentralized information, manipulated and filtered by a retinue of electronic agents.

A personalized newspaper serves as a good example. What would you want in such a newspaper? The question is tricky because, on the one hand, you would like stories that interest you. On the other hand, you would not entrust to any human or machine the wholesale dismissal of all other information. You might not normally be interested in news of a sporting event, but you know that your attitude would change if a close friend were competing. Newsworthiness is in the eye of the beholder—and not just the eye of the editor. Yet to date, the beholder has not been a part of the equation, except by coarse selection of one newspaper or magazine over another.

In the future, news and personal information systems will have the potential to be much more integrated than they are today, which should be highly beneficial. If your early-morning flight is delayed, that fact should appear as the lead headline in your personalized newspaper. Indeed, if the network is

"Tell Me More" Television


SEQUENTIAL TRANSMISSION OF A 16-HOUR VIDEO PROGRAM
TELEVISION RECEIVER STORES PROGRAMS IN MEMORY
5 SECONDS = 1 HOUR OF VIDEO
VIEWER BROWSES RANDOMLY THROUGH THE HOURS OF INTEREST

Packets of transmitted video information, all constituting one program, can be sent to a memory-equipped television receiver. The viewer can then browse through any of these hours of information in any order, seeking elaborations on subjects of interest or skipping what seems irrelevant. In this way, watching television can become more like reading a book or having a conversation.

smart enough, it should tell your alarm clock to let you sleep an additional hour (after checking for other airline routing options). And if your flight is on schedule, you should expect your morning edition to include the weather report for the city of your destination (so that you leave town with an umbrella in spite of the blue sky).

Some people cry at this point, "Tilt! I don't want my newspaper to be so personalized!" The arguments against the idea are usually twofold: "I count on the impersonal newspaper to be a common subject for conversation, and I assume other people know this as well," and "I don't want to become narcissistic and introverted by hearing only what I want to hear all the time."

These concerns are not totally misguided. Still, they assume that personalization is a full swing away from common knowledge, which it need not be. In fact, if you think of your best friends, they are people who do not tell you what you want to hear but are willing to tell you what they believe to be the truth and what is in your best interest. Such could be the case with an electronic agent as well; it need not be limited to what it thinks you want to know. The trick is timeliness. Our elasticity for less relevant information is very different on early Monday morning than it is on Sunday afternoon or when sitting in a garden compared with stepping over the threshold into an important meeting.

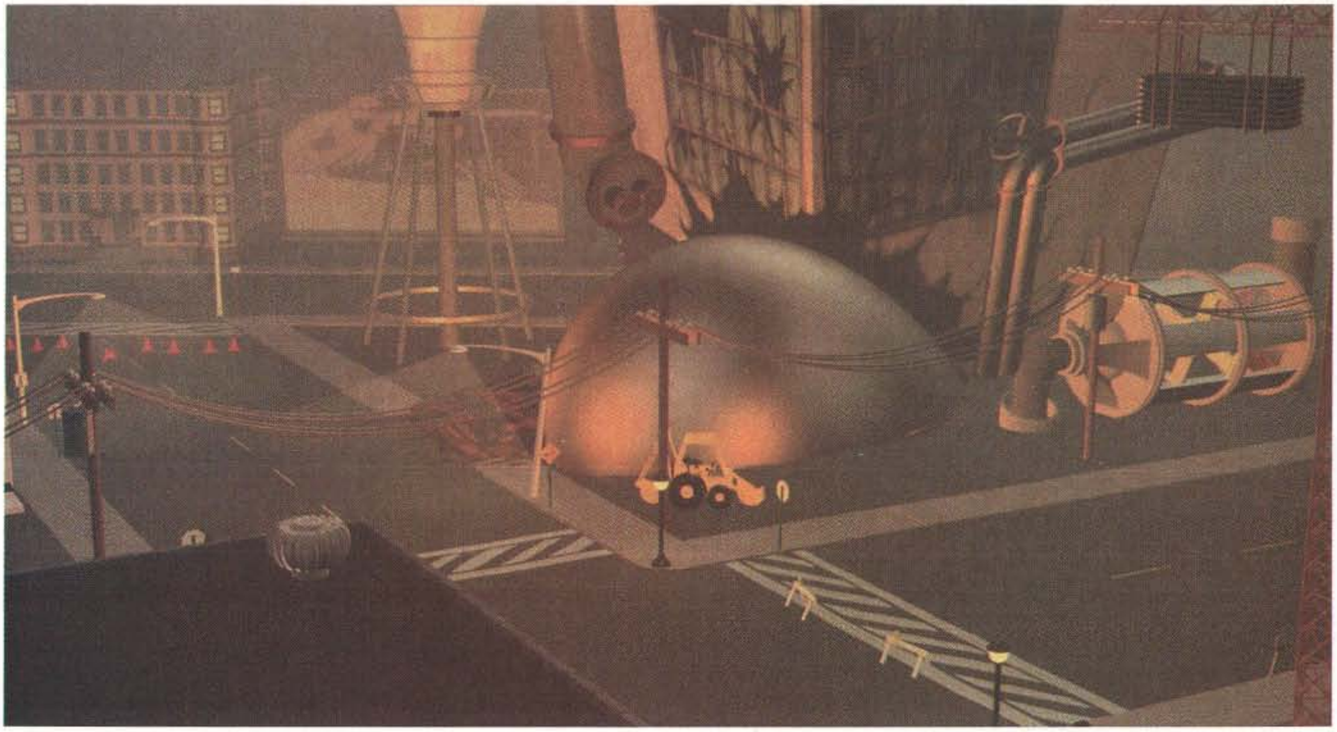
Consider personalization in the context of a telephone, which today rings indiscriminately. A well-trained, experienced butler will respond to a call on the basis of who is calling and, in many cases, of what the caller wishes to say. While this arrangement may sound one-sided, it is not. Often, the caller is delighted to get the butler and leave the message regardless of whether the intended recipient is there, just to ensure the timeliness of the message's delivery. The network of the future should behave no less intelligently.

A specific example of personalizing information can be found in the idea of individualizing television. The example is interesting in that it individualizes the premiere mass medium and, in some ways, turns the idea of television networks upside down.

In 1965 the Carnegie Corporation asked James R. Killian, Jr., former president of M.I.T., to chair a commission on educational television that laid the foundation for the Public Broadcasting System. As the commission approached the end of its work in 1967, the need for a more technical view of the future surfaced; J.C.R. Licklider of M.I.T., who was famous for his work on computer and human interfaces, was asked to provide a quick appendix. In it he coined the term "narrowcasting." He wisely foresaw an emerging need to make more specialized programs that would appeal to smaller audiences. That idea has been taken up enthusiastically by the magazine industry in particular and by the information distribution industry in general: they are doing more and more narrowcasting of all kinds. What Licklider missed was the once unthinkable limiting case: the audience of one.

Contrasting broadcasting with narrowcasting overemphasizes the difference between "broad" and "narrow." Instead we should be focusing on the word "cast." Stewart Brand, in his book *The Media Lab,* coined a term, "broadcatching," that draws attention rightly to the real challenge: to catch television and other media information selectively so that the sum of the collected parts is personalized.

As mentioned earlier, it is possible to transmit one hour of TV in five seconds. In principle, your television could look at more than 5,000 hours of programming during a seven-hour workday. One can imagine a system of television indexing. In the near future, programs could be digitally broadcast with a table of contents at the beginning. By consulting this table of contents, your intelligent TV set could filter out inappro-



THIS FRAME OF ANIMATION for high-definition television would require about 50 million bits to describe all the pixels; 30 frames per second are needed for animation. Transmitting that much information fast enough would take enormous bandwidth. If a computer in the receiver holds models of the pictured objects, however, only instructions about their movements need be sent. Smoothly moving high-resolution images are therefore possible with low bandwidths.

priate programs. It could then more or less edit the 5,000 hours of programming it has monitored into a short, distilled collage lasting no more than, say, 10 minutes. Elaborations on the contained subjects of interest could be obtained by instructing the TV to "tell me more." The total length of the stored programming would be limited only by the receiver's memory capacity. Broadcatching of this sort can be generalized to many media. The personalized aspect of the provided service would be handled by the receiver, not by the network as a whole.

Too often networks are likened to highways. By analogy, we are going from a system of country roads to one of autobahns, with wider lanes and higher (or no) speed limits. This comparison is misleading because it suggests that the network can stand by itself. That is not the case. The added intelligence at each node and at the ends of the network are what make the system work.

To understand the difference that intelligence makes, think about how printed information can be distributed electronically. A book could be stored in one location and shipped anywhere as a facsimile in no time at all. Each page would be fully encoded as a bit-map illustration, a faithful reproduction of the page, like a picture of it. At first glance, that prospect is tantalizing, especially to those of us who have suffered through 40- or 80-character columns of text while knowing that the printed originals had a rich, expressive and readable format. Yet on closer examination, storing and transmitting books as faxes is a dreadful mistake.

The right way to solve the problem is to store a description of a page format, including the typographics in which text is printed. That page description should be transmitted along with the ASCII code for the characters in the text and the bit-map information for the photographs and other structured graphics. At the receiving end, one needs the computer intelligence to decode the page description language and display the "pages" of the book.

Why is this approach so desirable? Two reasons. For one, the information in the book remains computer readable. Machines can refer to the text and decide whether to retrieve, route or file it. For another, in this form the information can be reformatted for display on a variety of media: you may occasionally want the information delivered to a display on your wrist or in your vest pocket rather than getting a facsimile of the original page.

PERSONALIZED ELECTRONIC NEWSPAPER can be compiled by computers scanning the network for information of particular relevance to specific individuals. A sophisticated electronic model of a person can be used to determine which news items should be selected and how much prominence and detail they should have.

Mind you, the mistake of facsimile is not limited to those people bent on finding excuses to consume bandwidth or promulgate dumb terminals. It is evident today in all our daily lives. The notion of "imaging" documents in general and the presence of fax machines in particular are setting back by decades the cause of making information more readable to computers.

Just think of how this inefficiency is part of your own behavior. You type a letter on a word processor, print it on a laser printer and stuff the hard copy into a fax (or use a fax modem). The last step in this chain just deleted the computer-readable format of the text; the recipient can no longer use the text in an information management system that would have access to the content. Yet we all do this because there are many more fax machines than coherent electronic mail systems (let alone systems capable of interpreting page description languages).

The fax, a dumb terminal par excellence, perfectly represents the services that result when we do not focus on the intelligence of the network and its ends but instead rely on the lowest common denominator of transceiver. Making such mistakes grossly limits the quality and originality of the products and services that can later arise. Let us not make them again with broadband networks; networks present real opportunities for liberating us that are much too enticing to forfeit.

FURTHER READING

THE HUMAN INTERFACE: WHERE PEOPLE AND COMPUTERS MEET. Richard A. Bolt. Van Nostrand Reinhold, 1984.

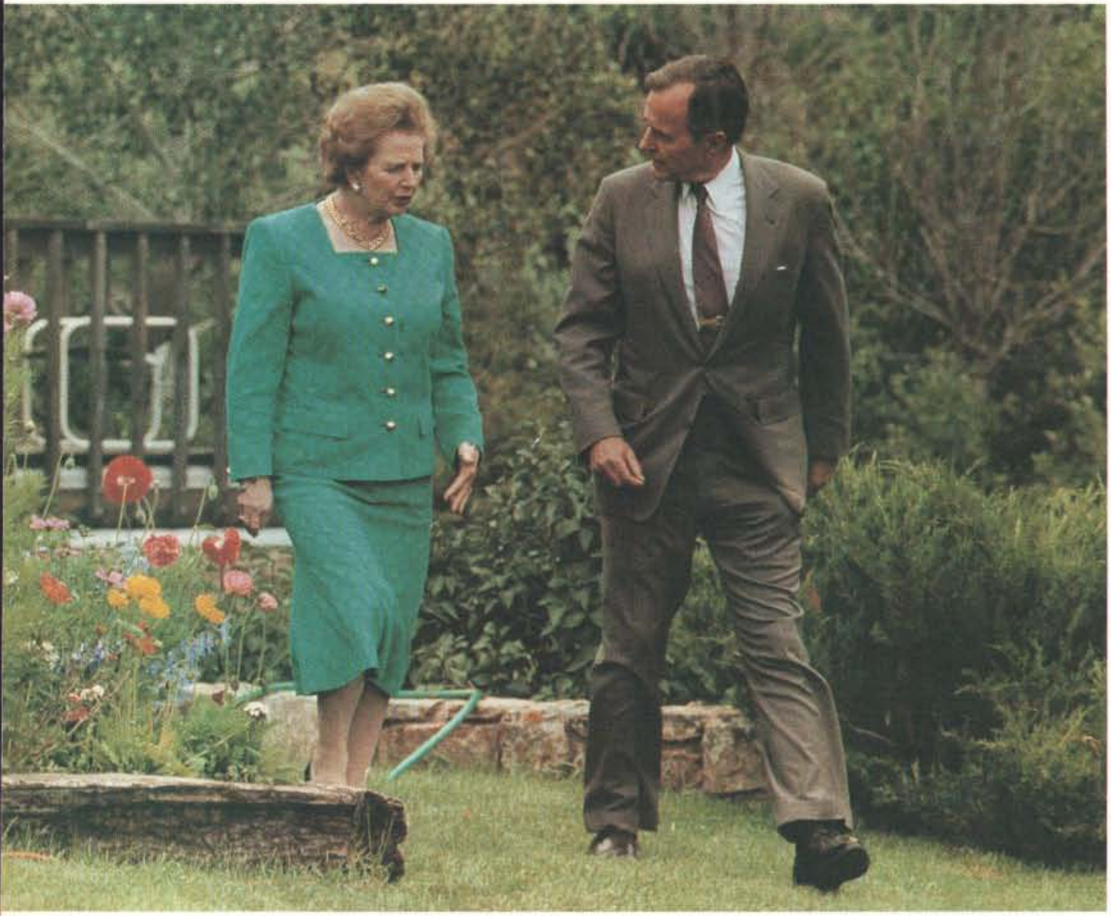
THE ART OF HUMAN-COMPUTER INTERFACE DESIGN. Edited by Brenda Laurel. Addison-Wesley Publishing, 1990.

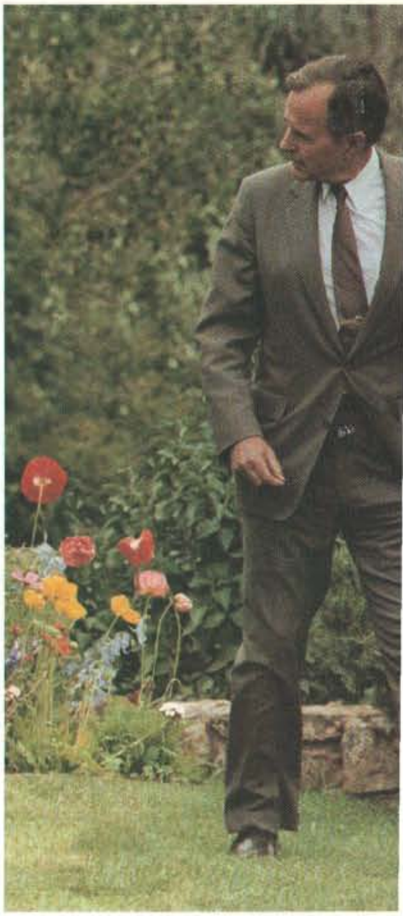
INTO THE TELECOSM. George Gilder in *Harvard Business Review*, Vol. 69, No. 2, pages 150–161; March/April 1991.

When Is Seeing Believing?

Digital technology for manipulating images has subverted the certainty of photographic evidence, so that a picture may mislead the viewer

by William J. Mitchell





The Manipulation Process

To create the "angry" image (*above, center*) from the original photograph (*above, left*), the computer artist first silhouetted the flowers out of the background. Next he cloned foliage from the upper right of the original and used it as a "paintbrush" to stipple greenery that replaces the porch behind the flowers in the original photograph. He softened, or blurred, Bush's outline and surround, a process called vignetting, and slid them under the silhouetted flowers. All the shrubbery behind the pair was blurred slightly.

The artist then silhouetted the rock in front of Thatcher and flipped it. The ferns that conceal the end of the hose in the

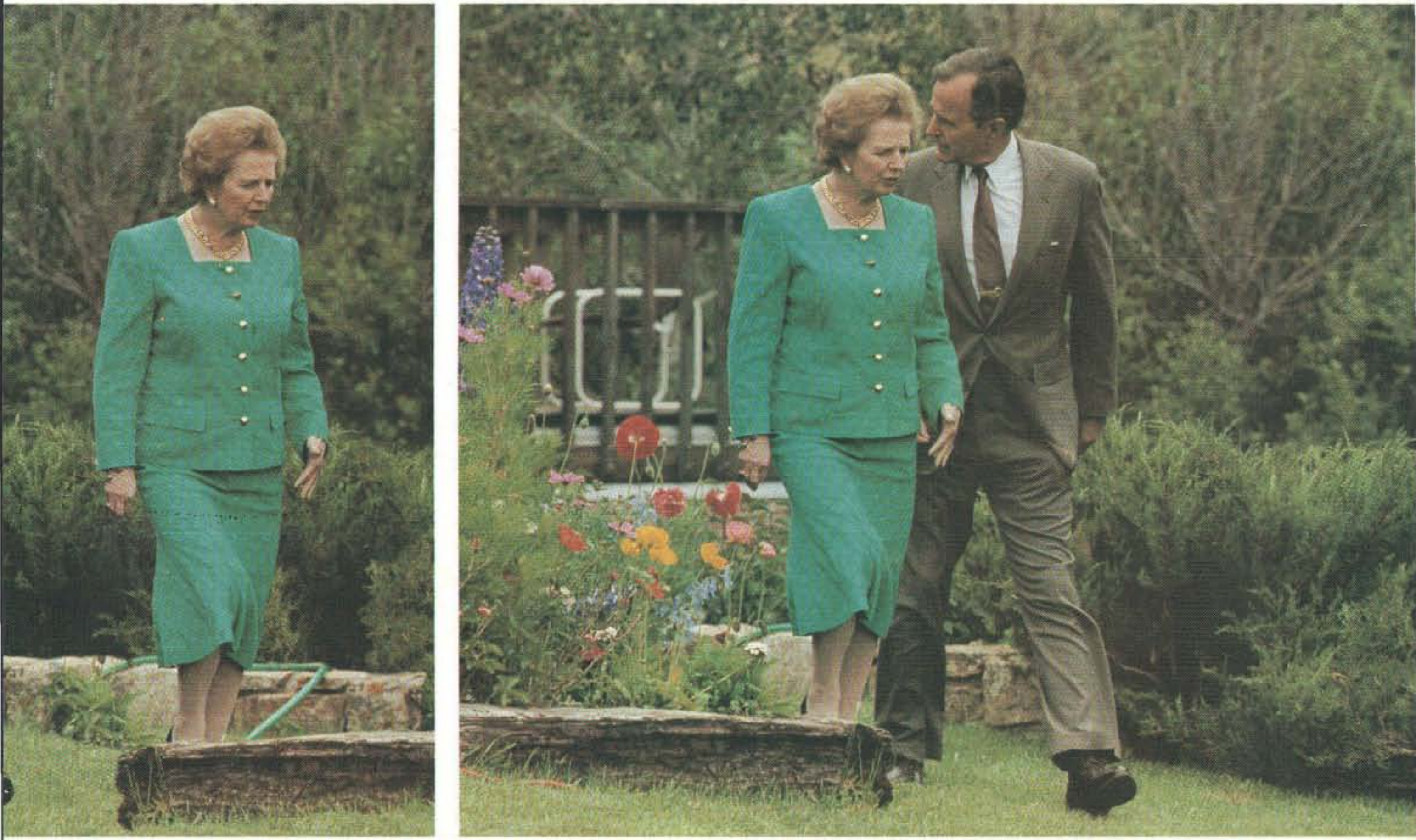
Reprinted from the February 1994 issue

In September 1993 newspapers around the world published an astonishing—almost unbelievable—photograph of Israeli prime minister Yitzhak Rabin shaking hands with Palestine Liberation Organization chairman Yasser Arafat on the White House lawn while President Bill Clinton looked on. In 1988 *Life* magazine published an equally striking picture of Chairman Arafat warmly greeting then Prime Minister Yitzhak Shamir under the approving gaze of President Ronald Reagan. One of these images recorded an actual event and provided reliable evidence that peace was perhaps at hand in the Middle East. The other was a computer-manipulated composite, a tongue-in-cheek fiction.

Such fake "photographs" can now be produced by using widely available "paint" and image-processing software to rearrange, recolor and otherwise transform the elements of a scene. The same software can combine fragments of different images into one new image. Other software can generate completely synthetic photorealistic pictures by applying sophisticated perspective projection and shading to digital models of three-dimensional scenes—a technique commonly employed to present architectural projects and to create Hollywood special effects.

Unlike drawings and paintings, which we regard as inherently untrustworthy products of human intention, these fakes can easily trick us into false beliefs. Like fingerprints

DIGITAL MANIPULATION of news photographs can become a novel form of spin doctoring. Two of these pictures were manipulated to suggest different emotions. What actually took place between George Bush and Margaret Thatcher—a chat, a quarrel or an intimate whisper?



original had to be duplicated and expanded to give the hose some place to go in the new image. And to the right of Bush's extended leg is a piece of the wall duplicated from that just to the right of the hose in the original. The new fern softens the transition between the two pieces of wall as well as hiding the hose.

In the "intimate whisper" image (*above, right*), Bush's vignetted surround was slid over the leafy shrub between him and Thatcher: Bush replaces bush.

Cut-and-Paste Photomontage

Traditional cut-and-paste methods transformed a 1923 photograph showing three Soviet farmers (*bottom*) into "evidence" that surfaced in July 1991 of the continued imprisonment of three lost fliers in Vietnam (*top*). Conventional photomontages, except when they are produced with great care by highly skilled artists, are relatively easy to detect as fakes. This one has many imperfections that give it away.

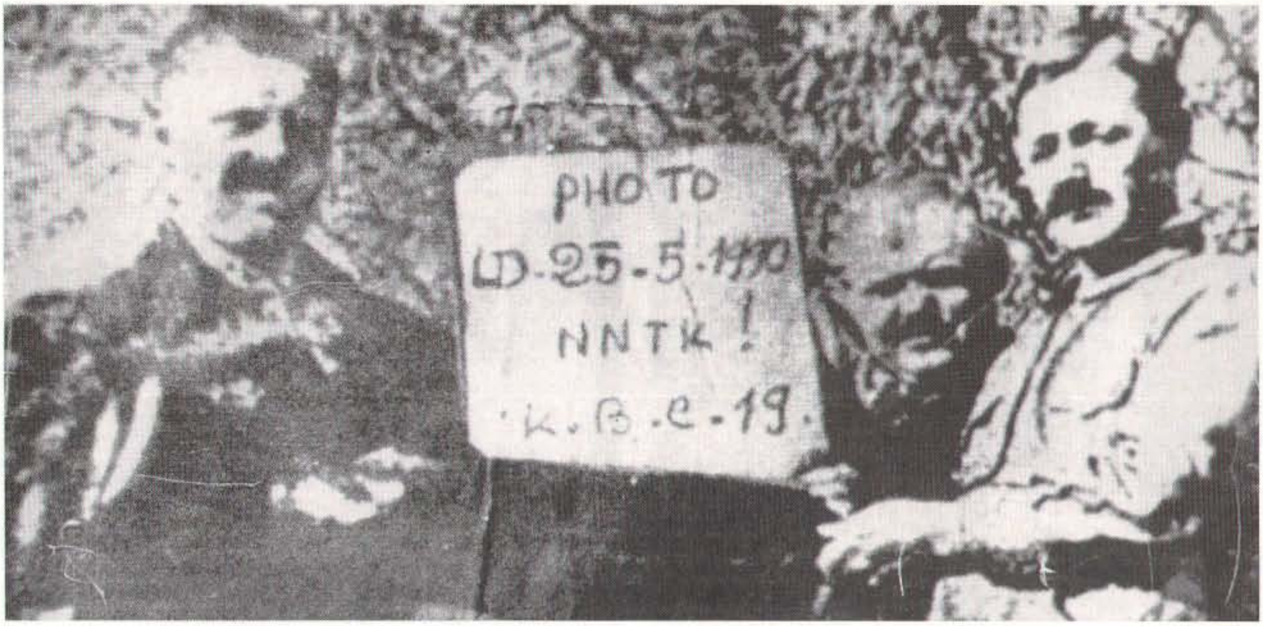
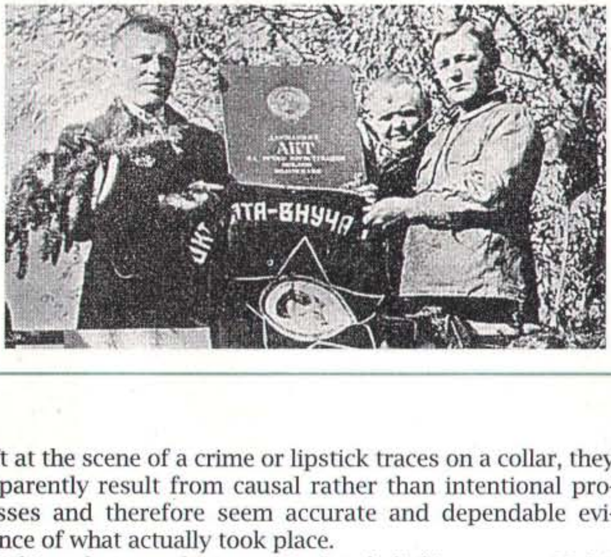


PHOTO
LD-25-5-1990
NNTK!
K.B.C.-19.



Look closely at the Stalinesque mustaches—especially the one on the left. Do they look fully convincing? Can you detect the handwork?

Should we read the sign as a foreshortened rectangle (with the right edge nearer the camera than the left) or as an irregular shape? Is it in front of the figure on the left, or is it behind? What is holding it up? (Note the hand at the top left in the original.) The spatial ambiguities suggest that it was pasted in.

Notice the fuzziness and poor tonal quality. This blurring may be deliberate, to conceal some of the imperfections. It also suggests generational loss resulting from rephotographing rather than printing from an original negative.

left at the scene of a crime or lipstick traces on a collar, they apparently result from causal rather than intentional processes and therefore seem accurate and dependable evidence of what actually took place.

Faking photographs to manipulate belief is not new. Multiple exposure and printing, cutting-and-pasting and retouching have frequently been used for hoaxes and for political propaganda. In the 19th century, for example, "spirit photographs" were produced by double exposure, and in the Stalinist era propagandists airbrushed the politically inconvenient Trotsky out of a famous photograph depicting Lenin addressing a crowd on May 5, 1920. But digital images make production of photographic falsehoods quicker and easier—and often much more difficult to trace. The question of how to distinguish visual fact from fiction is becoming increasingly urgent as we witness the explosive proliferation of digital-imaging technology. We are approaching the point at which most of the images that we see in our daily lives, and that form our understanding of the world, will have been digitally recorded, transmitted and processed.

Close inspection of traditional photomontages often reveals clear physical evidence of the doctoring that has taken place [*see illustration above*]. Printing masks and knife cuts may produce implausibly sharp edges; pencil marks and paint dabs may stand out against surrounding grainy textures; blends and reconstructions of surface textures may be imperfect; colors may not quite match. But digital images are manipulated by altering pixel values stored in computer memory rather than by mechanically altering surfaces. Therefore, digital photomontages typically show far fewer traces of the artist's hand. Furthermore, software for producing digital composites often simplifies the artist's task to such an extent that it may no longer take much time or much craftsmanship. Such software offers effective tools for tracing and blending edges, matching tones and colors and replicating textures. Thus, the fact that an image presents a seamless surface no longer provides strong grounds for concluding that it has not been manipulated. We must look for other kinds of clues—such as internal consistency, documentable provenance and consistency with existing beliefs.

"The faster you get there,

the more you can explore."

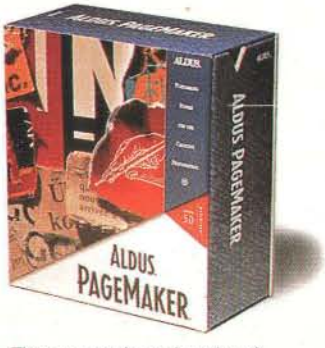
Power Macintosh

Tim Girvin
Tim Girvin Design
Seattle

"Design isn't just about creating a cool look anymore. It's also about solid business strategies." And about keeping 42 employees happy and productive. Not to mention keeping pace with clients like IBM, Nabisco, Nintendo and The Walt Disney Co.

"For me, the act of making **ALDUS PAGEMAKER** design is a spiritual thing–taking this energy, this idea, giving it life, and watching the vitality of a concept manifest itself in different ways. That's where native PageMaker on a Power Macintosh comes in. It's incredibly fast, so we can explore more manifestations in less time. And create better design. It's an enlightening experience. Profitable, too."

"When you're dealing with lots of complicated, multitiered documents that require constant reorganization, PageMaker and the Power Macintosh are a pretty incredible combination, unbeatable really."



The new, accelerated version of Aldus™ PageMaker™ combines its award-winning precision, flexibility and intuitive approach to document layout with the blistering speed of Power Macintosh. For more information and details on upgrades, call Adobe today at 800-685-3543.

Powered by the RISC-based PowerPC™ microprocessor, Power Macintosh™ takes desktop computing to new levels of performance, compatibility and productivity. For the name of the authorized Apple reseller nearest you, call Apple today at 800-732-3131, ext. 850.

Adobe

Apple

Accelerated software. The acceleration of discovery.

Because applications accelerated for Power Macintosh run so much faster, you have more time for creative exploration. Or another cup of joe.

Who needs faster software? Perhaps the better question is, "Who doesn't?"

Computational speed, after all, frees us to explore problems more deeply, allowing us to arrive at richer solutions. It frees us to relax our overworked synapses, resulting (if luck is with us) in new solutions that no number of hours at the keyboard could ever have produced. If we're disciplined, speed can even give us more time to pursue interests away from the office, the campus or the cramped home study.

Perhaps that's why so many people – from home and business users to software developers and scientists – are embracing Power Macintosh computers and the new generation of accelerated applications they have ushered in.

A path to productivity.

Because accelerated (or "native") applications have been optimized for the RISC-based PowerPC microprocessor at the heart of Power Macintosh, they run two to eight times faster than their non-native brethren. Which means you can complete every task you tackle that much faster, and move on to the dozens more awaiting your attention. It also means you have more time to refine all your work, to apply new concepts, to test new options. So everything – including you – ends up looking better than ever before.

Look a little further into the new optimized applications, and you'll see a lot more than raw speed. You'll also see new features and functions such as intelligent help, videoconferencing, speech recognition, text-to-speech conversion, desktop video and animation – things that simply weren't practical before the powerful combination of RISC technology and Macintosh® ease of use.

A path to compatibility.

Applications optimized for Power Macintosh run faster. And they do things you've never seen software do before. But what about all that perfectly functional software you already have running on your existing Mac™?

Don't worry. Power Macintosh runs it too. In fact, it runs *thousands* of existing programs for 68040 and 68030 Macintosh computers. For that matter, with SoftWindows™ software from Insignia Solutions, your Power Macintosh can run thousands of programs for DOS and Windows systems,* too.

A groundswell becomes a tsunami.

Long before we introduced Power Macintosh in March 1994, we were working hand in hand with software developers worldwide to ensure that a wide range of optimized applications would quickly become available.

And hundreds now are. From general-purpose business software from companies such as Microsoft, Claris, WordPerfect and Aldus to powerhouse specialty packages from companies such as Ray Dream, Wolfram Research, Macromedia, Virtus and more.

To get a better idea of applications available today, take a look at the box to the right. To get the latest, comprehensive list of *every* accelerated app by fax, or to learn about the exciting new software on the horizon, call us today at 800-770-4852. Or contact us via the Internet at http://www.info.apple.com.

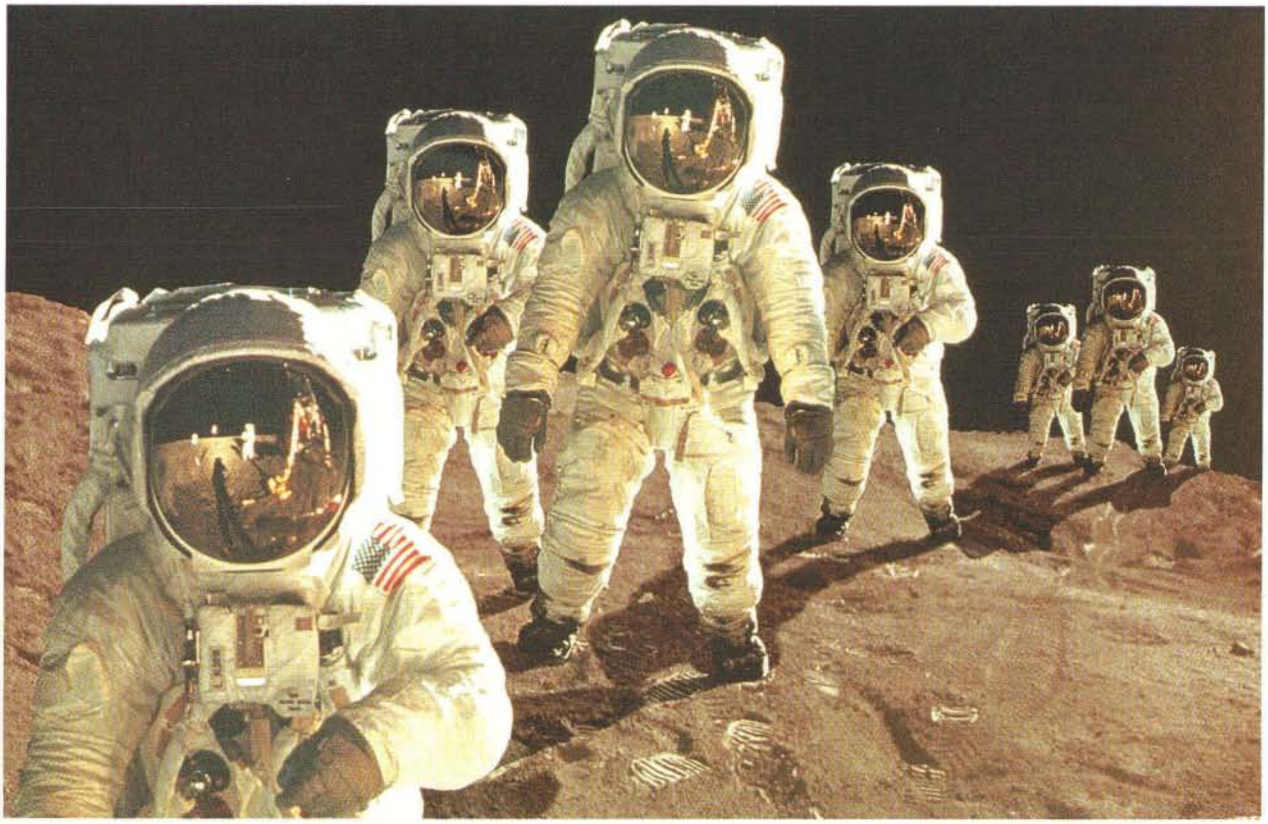
Adobe Photoshop
Aldus Persuasion
ArchiCAD (GraphiSoft)
Claris MacWrite Pro
MacFID (TECMAG)
Mathematica (Wolfram Research)
Microsoft Excel
Sketch! (Alias Research)
StatView (Abacus Concepts)
Virtual Gibbs
Virtus VR
WordPerfect

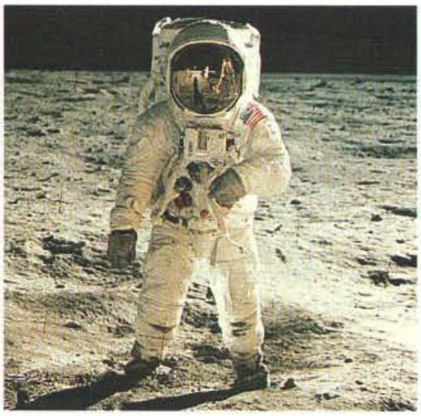
Apple

The Power Macintosh 6100/66 DOS Compatible runs DOS, Windows and Mac OS. Other Power Macintosh models require SoftWindows by Insignia Solutions to run DOS and Windows systems, and SoftWindows may need to be purchased separately. ©1995 Apple Computer, Inc. All rights reserved. Apple, the Apple logo and Macintosh are registered trademarks of Apple Computer, Inc. Mac and Power Macintosh are trademarks of Apple Computer, Inc. PowerPC is a trademark of International Business Machines Corporation, used under license therefrom. Adobe and the Adobe logo are trademarks of Adobe Systems Incorporated; Aldus and PageMaker are trademarks of Aldus Corporation, a subsidiary of Adobe Systems Incorporated. Windows is a trademark of Microsoft Corporation, and SoftWindows is a trademark used under license by Insignia from Microsoft Corporation. All other product names and logos are trademarks or registered trademarks of their respective companies. This ad was created using Aldus PageMaker 5.0 on the Power Macintosh.

Digital Photomontage

Digital photomontage of seven astronauts on the surface of the moon was produced from an original photograph made in 1969 by NASA of a single astronaut, Edwin F. Aldrin, Jr. (*bottom*). This montage is of high technical quality; it is carefully contrived to seem spatially consistent, and sophisticated digital technology has eliminated any obvious signs of cutting and pasting. So how can we tell it is a fake?





There is an obvious internal inconsistency. Because the composition was produced from scaled-down replicas of the original figure, the reflections in the visors are incorrect. Each shows the image of just one other astronaut, not the several we would expect.

There is implausible repetition—good grounds for suspecting the image was produced by replication operations. Is it likely that all the astronauts except one would be holding their left arms in precisely the same position?

There are some questionable cast shadows. The artist had to insert these with a digital "paintbrush" and faced the difficult task of making them consistent with the rough terrain. Do those at the back stand up to close, critical inspection?

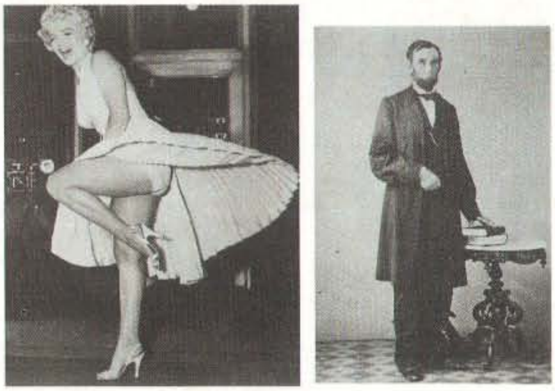
Such flaws and inconsistencies are very obvious once they have been pointed out, but they often go unnoticed at first glance. A skilled forger attempts to anticipate the types of cross-checking that a suspicious viewer will perform, then adjusts the visual evidence accordingly.

Freedom from internal inconsistencies does not demonstrate the veracity of an image. The existence of such flaws may, however, serve to refute claims that an image is a photographic transcription of physical reality [*see illustration above*]. Here are a few of the more obvious questions to ask when looking for inconsistencies. Do all the objects in the scene seem to be in correct perspective? Does the perspective foreshortening of a surface seem consistent with the spatial orientation suggested by its shading? Do indicators of time, such as clocks and shadows, all show the same moment of exposure? Do some objects seem surprisingly light or dark in relation to their surroundings? Are inserted objects betrayed by lack of shadows or by shadows cast at angles different from those cast by other objects? Are there shadows that do not seem to be cast by any object? Are shadows and specular highlights consistent with the assumptions about locations of light sources? Do unexpected discontinuities in the background of the scene suggest that objects have been deleted from the foreground? Do shiny surfaces display the expected reflections of other parts of the scene? Are surface intensities appropriately modified by diffuse interreflection effects? Are there plausible texture

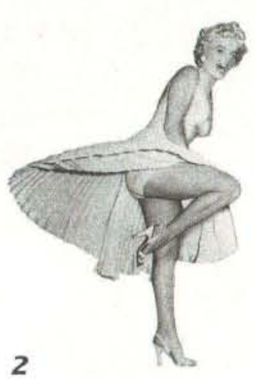
Creating a Digital Photomontage

For base images, the computer artist chose a publicity still of Marilyn Monroe made to promote *The Seven-Year Itch* in 1955 and a formal portrait taken of Abraham Lincoln in 1863 (*1*). To yield a richer final image, he scanned these black-and-white pictures into the computer as if they were in color. Once he had the images digitized, he flipped the Monroe photograph across its vertical axis and silhouetted her out of the background (*2*). In the Lincoln image, he extended the left side of the floor (*3*); the piece shown floating in space was copied and duplicated several times to provide the additional pattern needed.

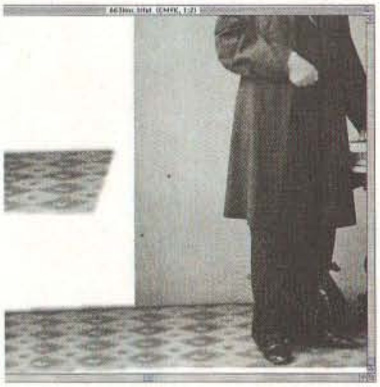
1



2

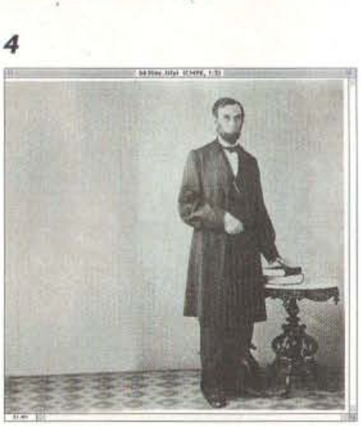


3

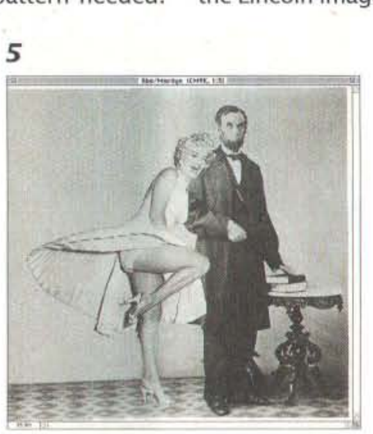


Next he filled the wall in roughly (*4*) in preparation for the silhouetted figure of Monroe, which he placed on top of the Lincoln image (*5*). He masked the two figures from the surrounding area so that the background could be evened out without affecting them (*6*).

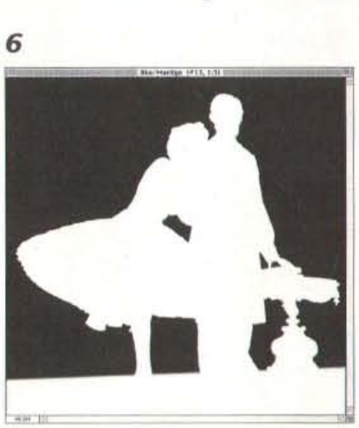
4



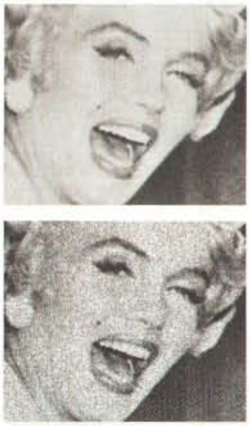
5



6

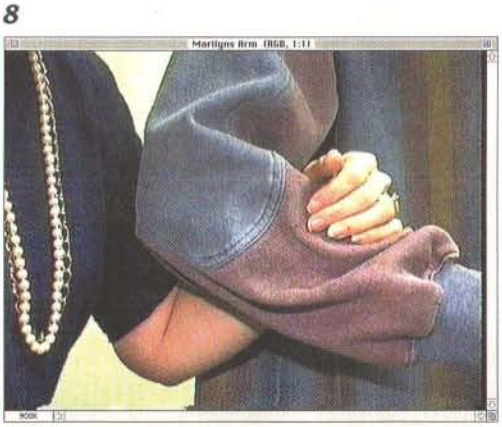


Close-ups (*7*) show Monroe before and after "noise" was added to degrade her image to match that of the older Lincoln photograph. (The noise function, which is built into the software, adds random pixels.) To provide Monroe with a hand and arm that she could tuck under Lincoln's elbow, the artist made a video of his wife holding the arm of his assistant (*8*), using lighting that simulated that in the original photographs. He then froze a frame of the video and captured it in his computer. The video was converted from color to gray scale and composited into the Lincoln-Monroe image (*9*).

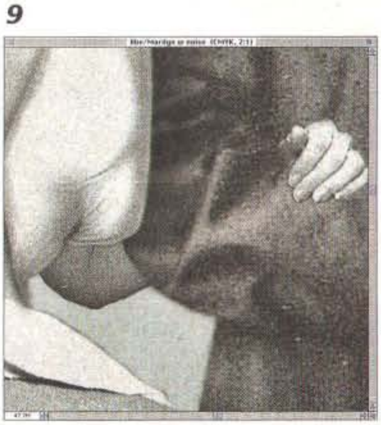


7

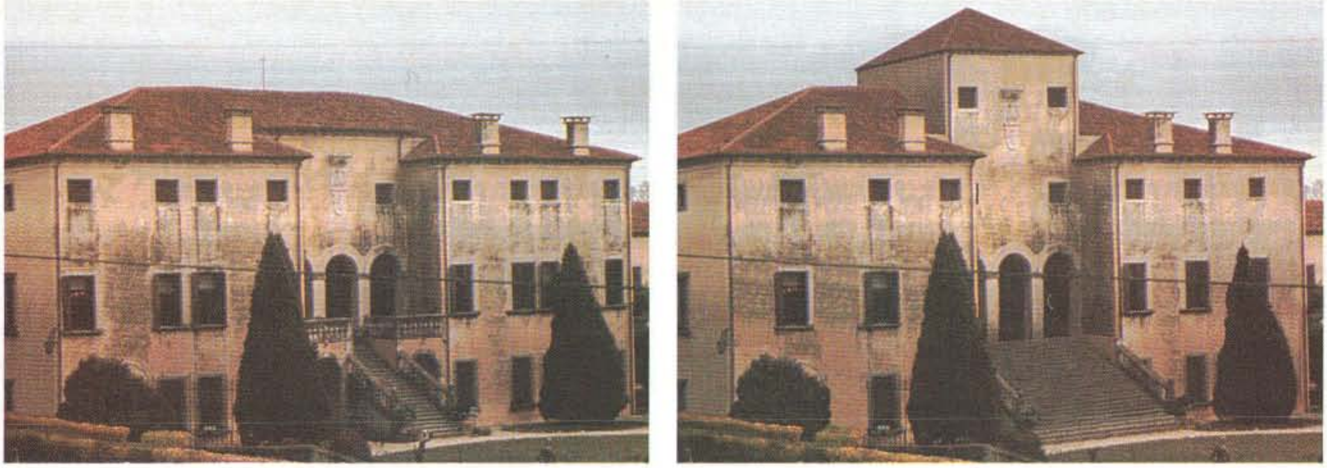
8



9



COMPUTER ART BY JACK HARRIS/ *VISUAL LOGIC*



WHICH VERSION IS A TRUTHFUL DEPICTION? Most people have no way to judge which of these two images produced for low-resolution video portrays the actual building. No internal clues reveal the fake. Scholars of Italian Renaissance architecture, however, can immediately recognize that the version with the tower (*right*) is Andrea Palladio's published but unbuilt scheme for the Villa Godi, whereas the version without a tower (*left*) shows the design that was actually constructed. The image of the unbuilt project was produced by shifting windows and chimneys, extending the narrow steps and replicating fragments of wall and roof surface to delineate the tower. Increasingly, our capacity to sort visual facts from falsehoods will rest on our ability to cross-check the visual evidence against established knowledge and beliefs.

gradients as surfaces recede? Do geometric perspective (size diminution with depth into the scene) and atmospheric perspective (color shift with depth) convey the same depth information? Is there a consistent gradient of sharpness from some focus plane?

In general, the more information there is in an image, the more difficult it is to change without introducing detectable inconsistencies. It is, for example, much easier to alter a fuzzy, low-resolution, dimly lit, black-and-white image than it is to do the same with a sharp, high-resolution, full-color picture. A photographic manipulator, like a dissembler who weaves a tangled web of lies and eventually trips himself up, runs the risk of being caught out by some subtle inconsistency that shows up when the visual evidence is carefully cross-checked.

Fully synthetic images produced by three-dimensional visualization software may, however, be free of such defects. Carefully modeled three-dimensional scenes, rendered in shaded perspective by ray tracing or radiosity (which calculates the interreflection of light within a scene), can simulate all the complex effects of light and shade that even the most demanding observer expects to see. Some fictional "photographs" may readily pass the internal consistency test.

We must then turn to such other criteria as the provenance of the picture. Because a photograph is exposed at a specific time and place, we can always ask to hear the story of how the photographer came to be at that very spot at that exact moment. Furthermore, we can ask for an account of how the picture subsequently made its way from the point of exposure to its present location. Sometimes the general credibility of a source—whether, for example, a purported picture of the surface of Mars is published in *Nature* or in a supermarket tabloid—can stand as a surrogate for an explicit narrative.

The provenance of a traditional photograph is often easy to trace, because exposed films, negatives and prints must be carried physically from place to place and because developing and printing must be performed in suitably equipped darkrooms. Digital-imaging technology makes this job much tougher: it eliminates negatives, it can replicate files in seconds, and digital images can be transmitted rapidly and invisibly through computer and telephone networks.

The subtlest challenge is posed by images that show no detectable signs of tampering and no obvious internal inconsistencies and yet contradict our established beliefs. Let us say, for example, we do not believe Elvis Presley still lives but are presented by an apparently reputable source with a sharp, detailed "photograph" of him in a recognizably contemporary setting. We can either maintain our confidence in the reliability of prior evidence that Elvis is dead and reject the image as a visual falsehood, or we can accept the new evidence before us and correspondingly modify our beliefs. Such intellectual judgments will be increasingly crucial and increasingly difficult to make with confidence in a world where convincing visual evidence can be faked with ease.

For a century and a half, photographic evidence has seemed unassailably probative. Chemical photography's temporary standardization and stabilization of the process of image making effectively served the purposes of an era dominated by science, exploration and industrialization. Photographs appeared to be reliably manufactured commodities, readily distinguishable from other types of depictions. They were generally regarded as causally generated, truthful reports about things in the real world, unlike more traditionally handcrafted images, which seemed notoriously ambiguous and uncertain human constructions. The emergence of digital imaging has irrevocably subverted these certainties, forcing us all to adopt a far more wary and vigilant interpretive stance. The information superhighway will bring us a growing flood of visual information in digital format, but we will have to take great care to sift the facts from the fictions and the falsehoods.

WILLIAM J. MITCHELL is author of *The Reconfigured Eye: Visual Truth in the Post-Photographic Era* (MIT Press, 1992). He is professor of architecture and media arts and sciences and dean of the School of Architecture and Planning at the Massachusetts Institute of Technology. His research focuses on computational techniques and digital media in design and the visual arts. His previous publications include *The Logic of Architecture* (MIT Press, 1990) and, with Malcolm McCullough, *Digital Design Media* (Van Nostrand Reinhold, 1991).

Computers and Architecture

Advanced modeling and rendering algorithms allow designers and clients to "walk" through buildings and redesign them long before construction

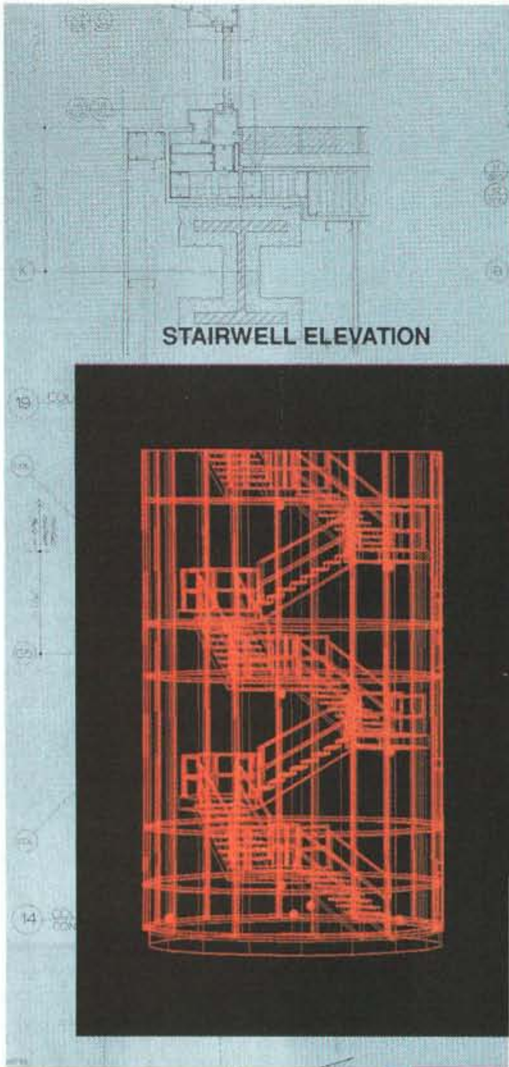
by Donald P. Greenberg

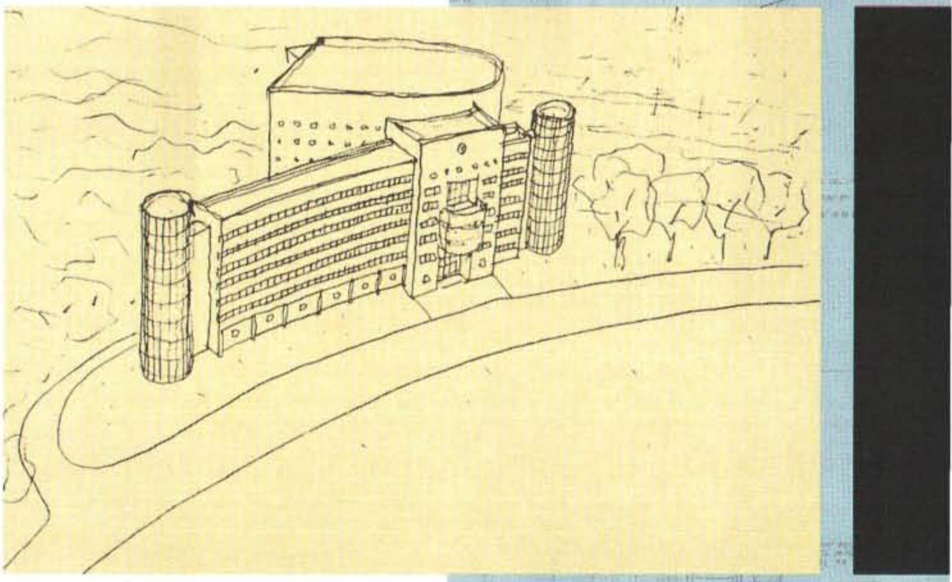
When it comes to designing a building and its interior spaces, architects still rely on sketches on the backs of napkins. They and their clients evaluate completed designs on the basis of hand drawings or painstakingly crafted miniature models. Many so-called computer-aided design systems employed by architects are in reality only computer aids to drafting. They automate the production of working drawings, but they do not provide the perceptual cues—color shading, shadows, texture and motion—that are necessary to evaluate an architectural space.

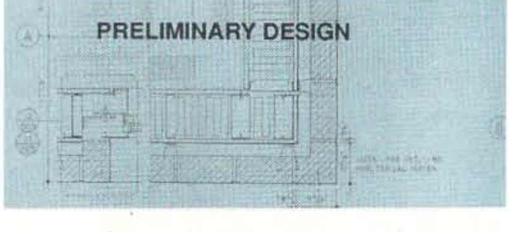
New software and advanced graphics-rendering techniques, however, are beginning to contribute to the creative side of architecture. These programs relieve the designer of labor-intensive drawing and provide more freedom to explore new ideas in three dimensions rather than just two. Architects can work through the entire preliminary design process—sketching and exploring aesthetic alternatives, refining their designs and generating realistic images for analysis.

Such sophisticated graphics programs, already at work in research laboratories and in architectural design courses at Cornell University, produce images much more useful to architects and clients than the drawings and models now employed. The images capture subtle nuances of light and shadow, and—most crucial to understanding how the finished design will work—they can be displayed in rapid sequence to give the impression of moving through a building as it will look after it has been constructed.

Architectural software must satisfy



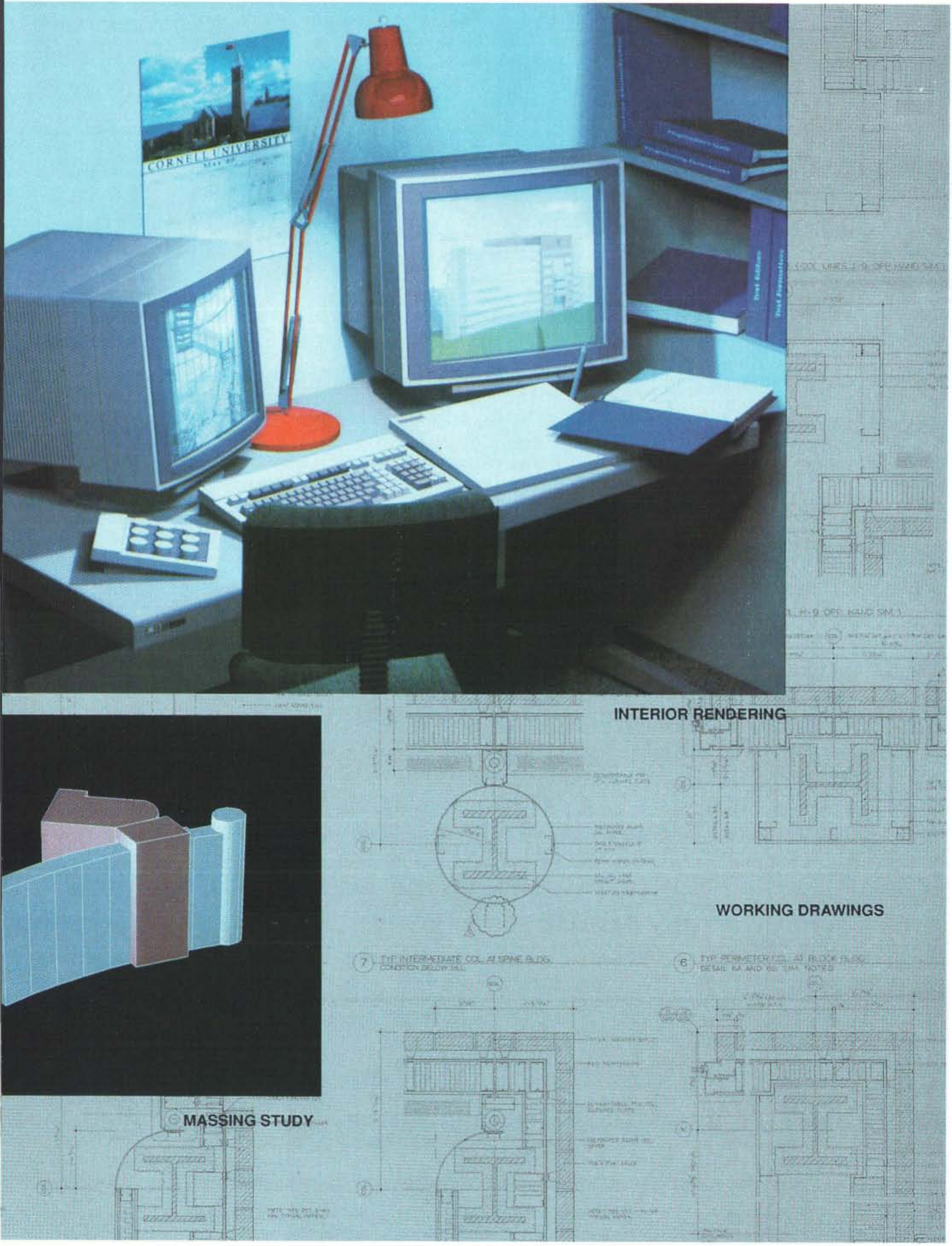




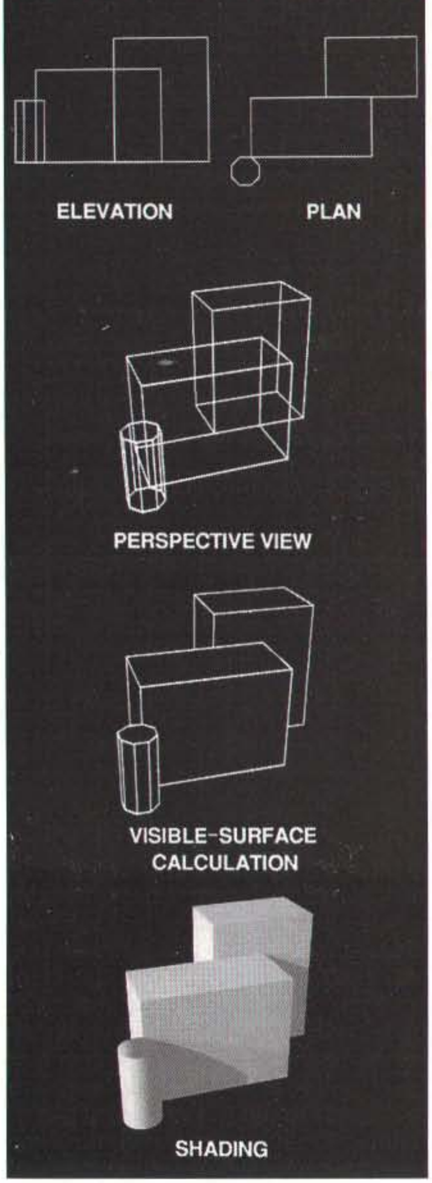
ARCHITECTURAL DESIGN process begins with simple sketches (*above*) and progresses through increasing levels of detail to the working drawings from which a structure (such as Cornell University's Theory Center) is built. New software and hardware can facilitate all but the earliest steps of the process.

DONALD P. GREENBERG is director of the computer graphics program at Cornell University, where he also teaches architecture. He and his colleagues apply computer graphics to modeling problems ranging from the design of buildings and stage sets and lighting to simulation of basic physics. He received his Ph.D. in structural engineering from Cornell in 1968.

Reprinted from the February 1991 issue


CORNELL UNIVERSITY
INTERIOR RENDERING
WORKING DRAWINGS
MASSING STUDY
7 TYP INTERMEDIATE COL. AT SPINE BLDG.
CONDITION BELOW SILL
6 TYP PERIMETER COL. AT BLOCK BLDG.
DETAIL 6A AND 6B SIM. NOTED

THE IMAGE-GENERATION PIPELINE



the unique characteristics and quirks of the architectural design process. Most important, design is iterative: if architects encounter a roadblock while refining a design, they must go back to a previous stage and rethink their initial decisions. During each stage, the mode of communication changes subtly. The initial process starts with doodles on tracing paper or sketches on the backs of envelopes. Later stages add more and more detail and refinement as areas and dimensions are fixed. Pencil gives way to ink. Standard details such as windows, doors and wall sections are incorporated in the drawings.

The preliminary stage of design focuses on the size and scale of a building and its relationships to existing structures nearby. Studies establish the basic shape and composition of the structure and provide such requirements as the number of square feet needed in different parts of the building. The next phase, design development, encompasses the refinement of plans, elevations and cross sections, traffic patterns, the clustering of private and public spaces, and the layout of individual rooms. The working-drawing phase—currently the most expensive and time-consuming—produces a set of detailed drawings that provide exact dimensions and material specifications so that the building can be constructed. They also ensure compliance with building codes and budget requirements, and they help to protect the architect against contractual misunderstandings.

Currently available commercial programs, aimed at producing working drawings efficiently, are far too unforgiving to support the iterative design process. The programs require too much information in too precise a form to be useful for doodling. Furthermore, drafting software is generally not designed to produce high-quality renderings suitable for communicating the feel of a building's spaces.

Historically that task has fallen to painstakingly constructed small-scale physical models. Such models are expensive and inflexible, and they can generally only provide a bird's-eye view of a building or interior. They certainly cannot give an architect or client a true sense of the perspective, colors, textures and shading that would be perceived by someone walking through the final structure.

Only computer graphics can do that. The extraordinary software and hardware advances of the past decade make it possible to generate full-color images of complex scenes rapidly enough to portray motion. Such modeling software can create renderings of complex environments containing specific materials with particular colors and textures—concrete, steel, glass and fabrics. Image-synthesis algorithms are now advanced enough to model even subtle nuances of light and shadow. Powerful workstations incorporating these algorithms permit architects to see the results of their design decisions immediately and to revise their designs interactively.

Creating a picture of a scene is a five-step process. First the architect must define the objects in it, their shapes, positions, orientations, material characteristics and surface finishes. The second step is to pick a point of view, a process analogous to positioning a camera and selecting and focusing the lens. The computer then transforms the object data to create a perspective image.

The third step is to determine which surfaces are visible to the observer. In the real world, physical laws take care of this problem, but computer-generated scenes require time-consuming calculations to solve the so-called hidden-surface problem. All objects in the scene must be checked to determine whether they overlap; if they do, only the object closest to the observer is displayed.

The fourth step is to simulate lighting. The light incident on any surface in the scene is a combination of that coming directly from sources (whose positions have been determined along with those of other objects in the scene) and light reflected from other surfaces (indirect lighting). The intensity of light reaching the observer, furthermore, depends both on the incident light and on the reflective qualities of each surface. To simplify the rendering process, many systems calculate only the direct illumination of surfaces.

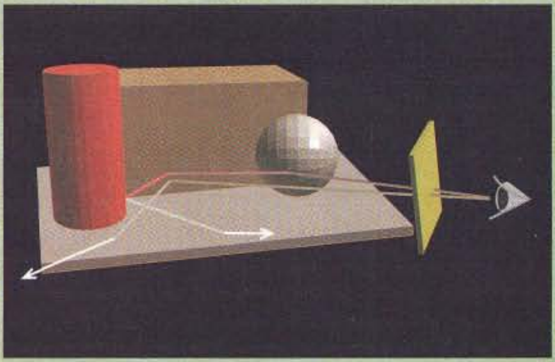
The last step is display. The rendering software must determine the amount of light that would reach the observer's eye from the real scene and generate an image that creates the same visual effect. The colors and intensities the software chooses may differ depending on whether the image is to be displayed on a screen, printed or shown as a color slide.

In modern high-performance graphics workstations, special-purpose hardware directly implements perspective transformations, visible-surface determination and display routines. This hardware can also execute the algorithms required to model simple direct illumination. Workstations carry out the graphics computations fast enough so that architects can model their designs and evaluate them easily at early stages.

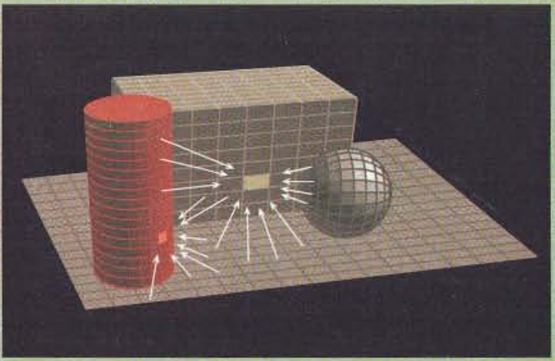
The quality of images produced by most workstations is limited, however. Pictures can easily be recognized as computer generated because they do not include the effects of interreflections among objects. As a result, they are useful primarily for initial modeling. Design evaluation requires much more realistic images, generated by more sophisticated software. Furthermore, these images must be displayed in rapid sequence to create the illusion of motion.

Computer graphics researchers have devoted considerable effort to the problem of lighting. The effects of lighting and reflection in real scenes are very

TWO METHODS FOR SIMULATING LIGHT AND SHADING



RAY-TRACING method for generating computer images sends imaginary rays of light from the viewer's eye into the scene being rendered. The color and light intensity at each point in the image depend on the reflectivity of the surfaces encountered by the ray in its passage through the scene.



RADIOSITY ALGORITHM calculates the light intensity of each surface in a scene as a function of the intensity of all other surfaces. Once the resulting set of simultaneous equations has been solved, the scene can be rendered rapidly from any angle.

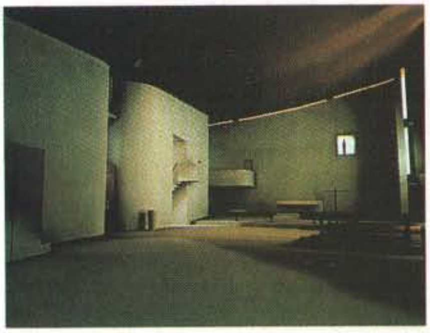
complicated and subtle. Indirect lighting is particularly difficult to model. It is possible to write equations that accurately simulate this global illumination, but they consume unreasonable amounts of processing time.

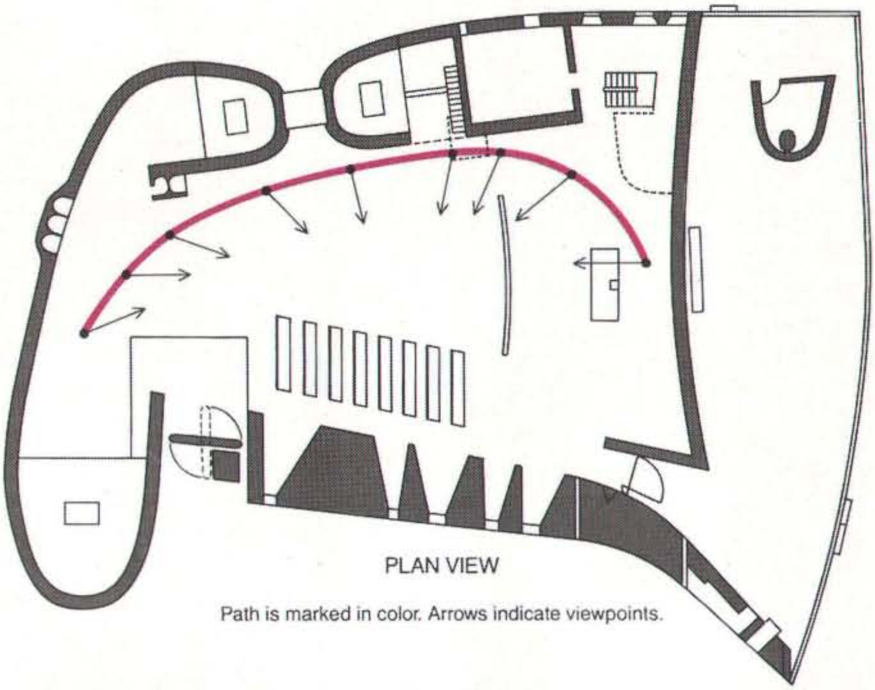
Two short-cut methods have become popular in the computer graphics community: ray tracing and radiosity. These techniques represent almost diametrically opposite approaches to generating an image. Ray tracing concerns itself only with light that enters the observer's eye and is particularly appropriate for environments that contain highly reflective surfaces. Radiosity, in contrast, determines the distribution of light energy throughout a scene and is best suited to environments containing mostly matte surfaces.

Although ray tracing produces extraordinarily realistic images and can certainly be used for architectural presentations, its value for design evaluation is limited. The approach is view dependent, and so the entire computation must be repeated for each new viewpoint. Ray-tracing algorithms work too slowly to generate dynamic sequences that show an architect or client what it would look like to walk through a space.

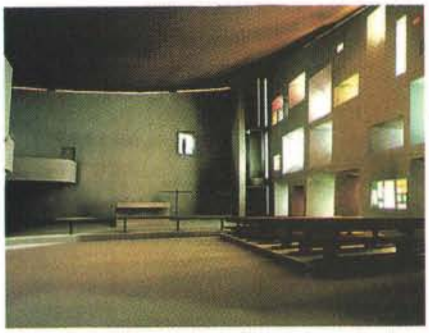
The radiosity approach, developed at Cornell in 1984, is view independent: it makes its calculations only once for a given scene. Once the global illumination has been determined, it is easy to create a series of pictures from different viewpoints. Indeed, the algorithm

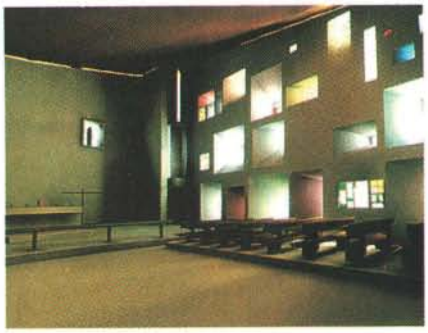
A WALK THROUGH LE CORBUSIER'S CHAPEL AT RONCHAMPS

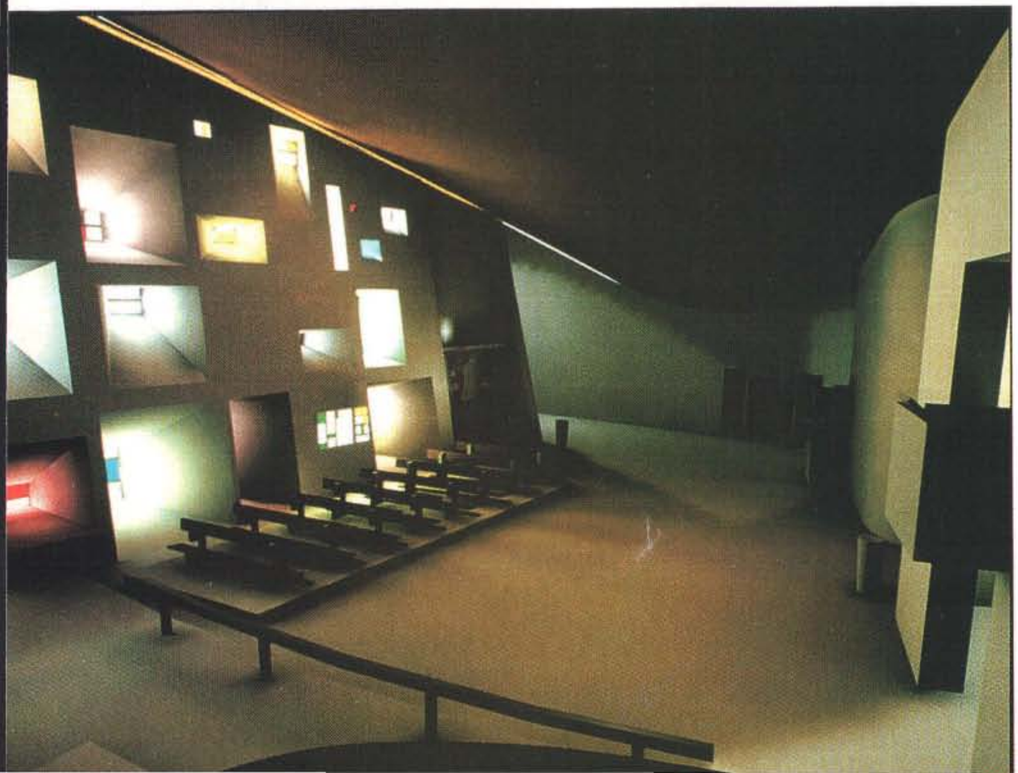




Path is marked in color. Arrows indicate viewpoints.





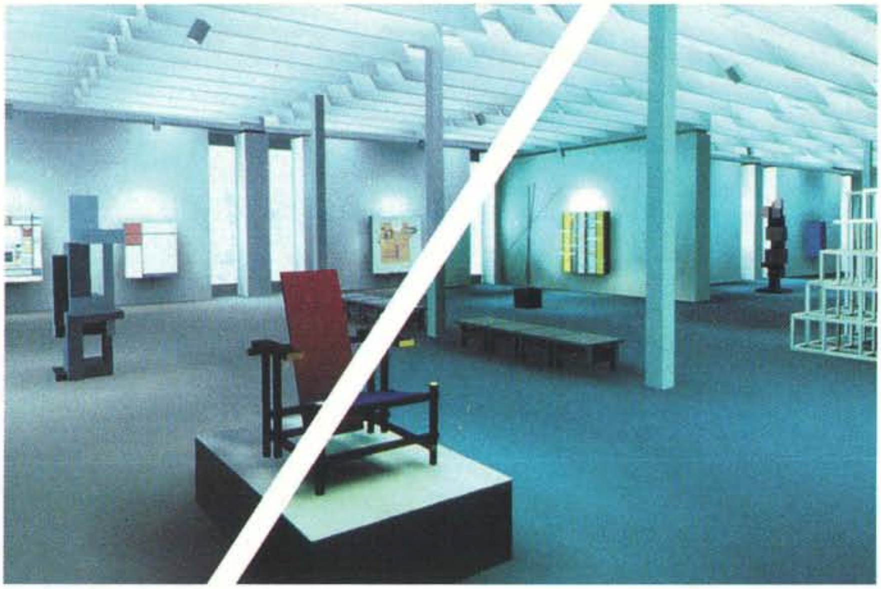


COMPUTER MODEL gives the impression of moving through a space. Image-rendering software running on high-performance graphics workstations can generate views of a structure from any position at rates of several frames per second or more. Such walk-throughs are of use both in evaluations of structures yet to be built and in studies of existing buildings. The balcony at Ronchamps (*left*), for example, is not usually accessible to the public.

can generate images fast enough to give the illusion of motion.

Ray tracing reverses the usual laws of light propagation. In the real world, light rays spread out from each light source in a scene and reflect from surfaces in the scene; eventually a small proportion of all the light rays pass into the eye of the observer. Instead ray-tracing algorithms follow the paths of the light rays back from the eye through the scene to the surfaces and light sources that engendered them.

The most popular method for modeling global illumination using ray tracing was introduced by Turner Whitted of Bell Laboratories in 1979. It traces the path of a single ray from the eye through each point in the image plane (each pixel on a workstation screen, for example) into the environment. Whenever the ray strikes a surface, it spawns reflected or refracted rays, which in turn are traced to establish the surfaces they intersect. The final intensity of each pixel is determined by adding up the light contributed by each spawned ray.



COLOR AND TEXTURE of computer renderings (such as this hypothetical museum of constructivist modern art) can be changed simply and quickly. Architects and their clients can evaluate different choices of construction materials and color schemes at minimal expense.

Radiosity relies on the conservation of energy to determine accurately the light intensity for each surface in a scene composed of ideal diffuse emitters and reflectors. An equation can be written defining the radiosity of each surface in a scene—the intensity of light leaving the surface—as a function of the radiosity of all the other surfaces. The radiosity of an object depends on two factors: emission (if the object is a light source) and reflection of incident light. The incident light, in turn, depends on emission from all the light sources in the environment and reflections from all the other surfaces.

The radiosity values for all the surfaces are then expressed as a matrix of simultaneous equations that can be solved to yield values for the illumination of any given part of the scene. Currently the computation is tractable only for diffuse environments; complex patterns of emission or reflection make the equations too difficult to solve quickly.

Radiosity is not only fast (after the initial calculations have been done); it is also very accurate. It can reproduce such phenomena as "color bleeding," in which reflections from a red surface, for example, might make an adjacent white surface look pink, variations of shading within shadows, and the fuzzy shadow edges produced by extended light sources such as skylights or fluorescent light fixtures.

Computer-aided design systems can now produce the kinds of images architects need. The most important issue yet to be resolved is software that makes it easy for architects to create their models. Researchers are only now coming to grips with the kinds of operations required to design a building—creating shapes, specifying surface qualities, joining objects together into seamless composites. Although the initial rough sketches for a design may always be done on paper, the coming decade will see programs that are simple to learn and to use yet powerful enough to handle the design of entire buildings down to the smallest detail and capable of generating moving, photorealistic images.

FURTHER READING

COMPUTER GRAPHICS IN ARCHITECTURE. Donald P. Greenberg in *Scientific American*, Vol. 230, No. 5, pages 98–106; May 1974.

PRINCIPLES OF INTERACTIVE COMPUTER GRAPHICS. Second edition. William M. Newman and Robert F. Sproull. McGraw-Hill, 1979.

FUNDAMENTALS OF INTERACTIVE COMPUTER GRAPHICS. James D. Foley and Andries Van Dam. Addison-Wesley, 1982.

Heads in the Cloud

Can programmers entice consumers into the net?

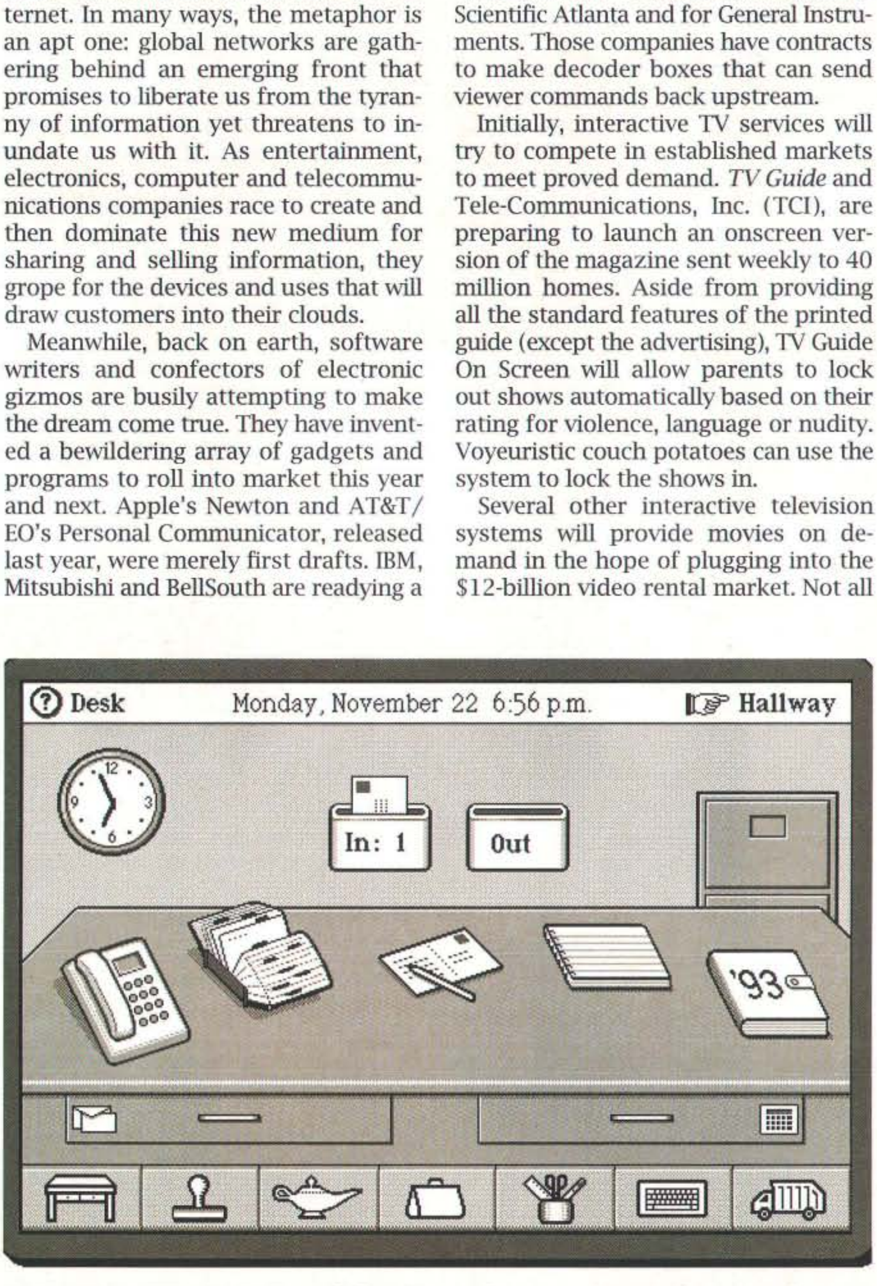
To the engineers who design digital superhighways, the symbol of a cloud represents far-flung, interconnected networks such as the Internet. In many ways, the metaphor is an apt one: global networks are gathering behind an emerging front that promises to liberate us from the tyranny of information yet threatens to inundate us with it. As entertainment, electronics, computer and telecommunications companies race to create and then dominate this new medium for sharing and selling information, they grope for the devices and uses that will draw customers into their clouds.

Meanwhile, back on earth, software writers and confectors of electronic gizmos are busily attempting to make the dream come true. They have invented a bewildering array of gadgets and programs to roll into market this year and next. Apple's Newton and AT&T/EO's Personal Communicator, released last year, were merely first drafts. IBM, Mitsubishi and BellSouth are readying a one-pound cellular phone/fax/modem with a computer in its handle. Motorola's entry is the first of a planned family of palmtop computers with built-in wireless fax, pager and electronic mail capability. Sony, Philips and Matsushita will follow with similar devices.

Hand-held communicators are but one enticement. Coming soon to a television near you are a raft of new computer-based entertainment and information services that will flow through the cable and telephone lines that already penetrate most American homes. Such "interactive TV" could be a bonanza for Scientific Atlanta and for General Instruments. Those companies have contracts to make decoder boxes that can send viewer commands back upstream.

Initially, interactive TV services will try to compete in established markets to meet proved demand. *TV Guide* and Tele-Communications, Inc. (TCI), are preparing to launch an onscreen version of the magazine sent weekly to 40 million homes. Aside from providing all the standard features of the printed guide (except the advertising), TV Guide On Screen will allow parents to lock out shows automatically based on their rating for violence, language or nudity. Voyeuristic couch potatoes can use the system to lock the shows in.

Several other interactive television systems will provide movies on demand in the hope of plugging into the $12-billion video rental market. Not all will use cable as a pipeline. Bell Atlantic is scaling up its StarGazer system for a field trial. StarGazer uses special transmission equipment and sophisticated video compression techniques to store digitized movies and mail-order catalogues on computer servers and pump them through ordinary telephone lines into homes.



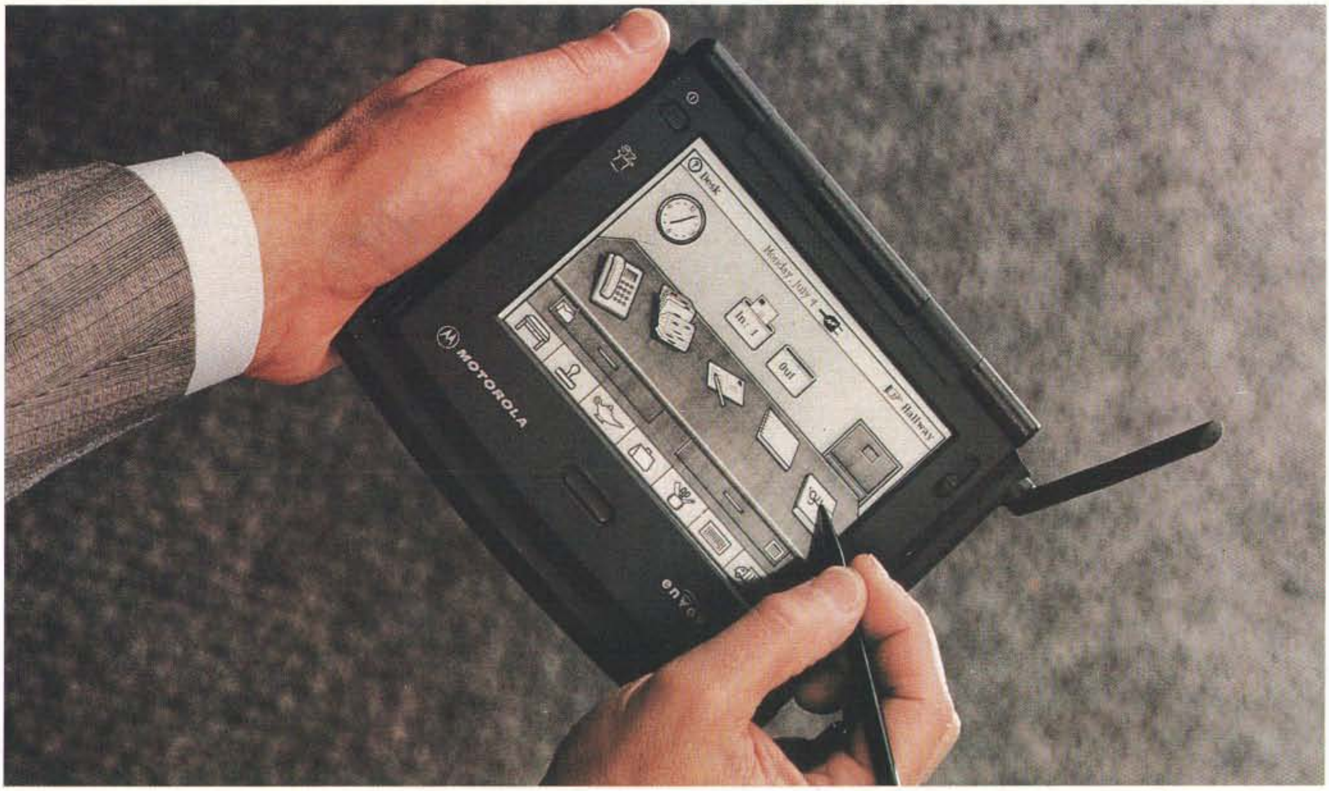
HAND-HELD COMPUTERS need friendlier software to bring cyberspace down to earth. General Magic's user interface gives applications a more mundane look.

Program guides, rental movies and shopping channels are just the first steps in a strategy to hoist a new industry by its bootstraps. "Most new media start off recapitulating the old media," observes Nathan P. Myhrvold, senior vice president for advanced technology at Microsoft (which has its own interactive television software). "Later on they start doing the really cool stuff."

Myhrvold's wisdom is as conventional as it is true. The competitor who establishes the dominant technology wins. Bell Atlantic, *TV Guide,* Time Warner and every other contender thus seek to build the pier at which the as yet unbuilt fleets of video information and entertainment services will come to dock. Bell Atlantic designed StarGazer from the beginning to be open to new services. As a common carrier, it must offer "video dial tone" service to any company that asks for it. When the system is completed—probably not until 1996—customers will be able to steer their television sets through a "virtual" shopping mall. "You might go into the EchoTV store to get the six o'clock news that you happened to miss or into the MGM store to select any MGM movie," muses spokesperson Missy McTamny.

"As much as we consume media, we spend more time communicating," Myhrvold remarks, pointing to another carrot with which the industry will entice consumers into the net. Motorola has already ridden to record profits on the fourfold increase in pager transmissions and 20-fold growth in cellular subscribers since 1985. This despite a general slowdown in consumer electronics sales that is estimated to have cut the profits of Sony, Sharp and Toshiba by 8 to 30 percent in 1993.

To computer executives, who face commoditization of their own industry, the path is clear: produce hand-held, computerized communications devices and software that actually makes them useful. Apple's Newton, designed as a "digital assistant," met with chilly reviews in the trade press, which criticized it for not focusing enough on communications. While Apple recoups, two other visions—that of the giant Microsoft and



NEW ENTRY in the hand-held communicator market is Motorola's Envoy. It has built-in facsimile, pager and electronic mail capability. The device offers a user two-way wireless and wire communication.

that of the tiny Silicon Valley start-up General Magic—will probably rise as top contenders in what is sure to be a messy struggle.

Almost everyone in the industry agrees that if hand-held devices are ever to capture the interest and dollars of the mass market, they must be substantially easier to use than desktop computers. Almost everyone but Myhrvold. "We've become inured to the fact that incredibly popular consumer devices are very hard to use," he says. "Consider a movie. You have to stand in line, it starts with or without you, and it's expensive. We go to movies because they make us laugh or cry, not because they are convenient. The trick is to make the ratio of how compelling something is to how hard it is to use greater than one."

Microsoft's idea of compelling starts in the workplace, so the company calls its new platform Microsoft At Work. "We thought the best thing is to put our software into fax machines and copiers and telephones—all in a way that integrates super-well with your desktop PC," Myhrvold explains. At Work also lies at the heart of WinPad, the operating system Microsoft is designing to run on devices built by Compaq and Motorola.

General Magic would like such evolution to look more like a revolution. It has decided to go straight to the consumer. Without benefit of Microsoft's incumbency or Goliath proportions, General Magic may seem an unlikely David. But a gang of big brothers—Apple, AT&T, Matsushita, Motorola, Philips and Sony—has invested in the company and licensed its technology. "We think of ourselves as a *keiretsu*," says Marc Porat, General Magic's chief executive officer and chairman.

General Magic hopes it can seduce consumers as well as gadgeteers by making its Magic Cap operating system more adept at communications and much friendlier than the product's competitors. "Abstraction is the enemy," says Andy Hertzfeld, whose official title at General Magic is Software Wizard. "Magic Cap goes further down the road of making computing tangible."

Whereas on a Macintosh, say, folders lie on a plane called the desktop and open to reveal a window containing files—badly mixed metaphors—Magic Cap tries to emulate wherever possible the way things look and work in the real world. The desktop *looks* like a desk, on which sit a telephone, a rolodex, a notepad and other items that do just what you would expect them to. Away from the desk one can wander around a corridor, opening doors to other applications. Or leave the building altogether and walk down a street on which sit various buildings, each representing an information service.

Making electronic mail more like the real thing poses a larger challenge. To most of the world, E-mail means text—and nothing else. So General Magic humbly plans to change the world. It has developed an interpreted programming language, called Telescript, that allows a message to include images, sounds and even instructions as to how it should behave. A message could schedule an airline flight, for instance, and then monitor the flight and notify you of any changes. Magic Cap has Telescript built in, and all Magic Cap devices (the first of which is the Motorola product) will come bundled with AT&T's MagicMail. But that is, so far, the only network that has licensed Telescript, although Porat promises that "we will license Telescript to anyone," even Magic Cap's competitors.

The competitors may have other ideas. "The notion that their programming language and the limited features it has will be the only way to do things is completely wrong," Myhrvold fumes. Although Telescript would make it much easier for programmers to develop the "really cool stuff" that will lure customers into this new marketplace among the clouds, Myhrvold hints that Microsoft will be proposing alternatives. "I've got 400 people and $100 million a year working on this stuff. It may not generate significant revenue for three to five years. But it sure is a hell of an opportunity." —*W. Wayt Gibbs*

Computers, Networks and Work

Electronic interactions differ significantly from face-to-face exchanges. As a result, computer networks will profoundly affect the structure of organizations and the conduct of work

by Lee Sproull and Sara Kiesler

Although the world may be evolving into a global village, most people still lead local lives at work. They spend the majority of their time in one physical location and talk predominantly to their immediate coworkers, clients and customers. They participate in only a few workplace groups: their primary work group, perhaps a committee or task force and possibly an informal social group.

Some people, however, already experience a far more cosmopolitan future because they work in organizations that have extensive computer networks. Such individuals can communicate with people around the world as easily as they talk with someone in the next office. They can hold involved group discussions about company policy, new product design, hiring plans or last night's ball game without ever meeting other group members.

The networked organization differs from the conventional workplace with respect to both time and space. Computer-based communication is extremely fast in comparison with telephone or postal services, denigrated as "snail mail" by electronic mail converts. People can send a message to the other side of the globe in minutes; each message can be directed to one person or to many people. Networks can also essentially make time stand still. Electronic messages can be held indefinitely in computer memory. People can read or reread their messages at any time, copy them, change them or forward them.

Managers are often attracted to networks by the promise of faster communication and greater efficiency. In our view, the real potential of network communication has less to do with such matters than with influencing the overall work environment and the capabilities of employees. Managers can use networks to foster new kinds of task structures and reporting relationships. They can use networks to change the conventional patterns of who talks to whom and who knows what.

The capabilities that accompany networks raise significant questions for managers and for social scientists studying work organizations. Can people really work closely with one another when their only contact is through a computer? If employees interact through telecommuting, teleconferencing and electronic group discussions, what holds the organization together? Networking permits almost unlimited access to data and to other people. Where will management draw the line on freedom of access? What will the organization of the future look like?

We and various colleagues are working to understand how computer networks can affect the nature of work and relationships between managers and employees. What we are learning may help people to exploit better the opportunities that networks offer and to avoid or mitigate the potential pitfalls of networked organizations.

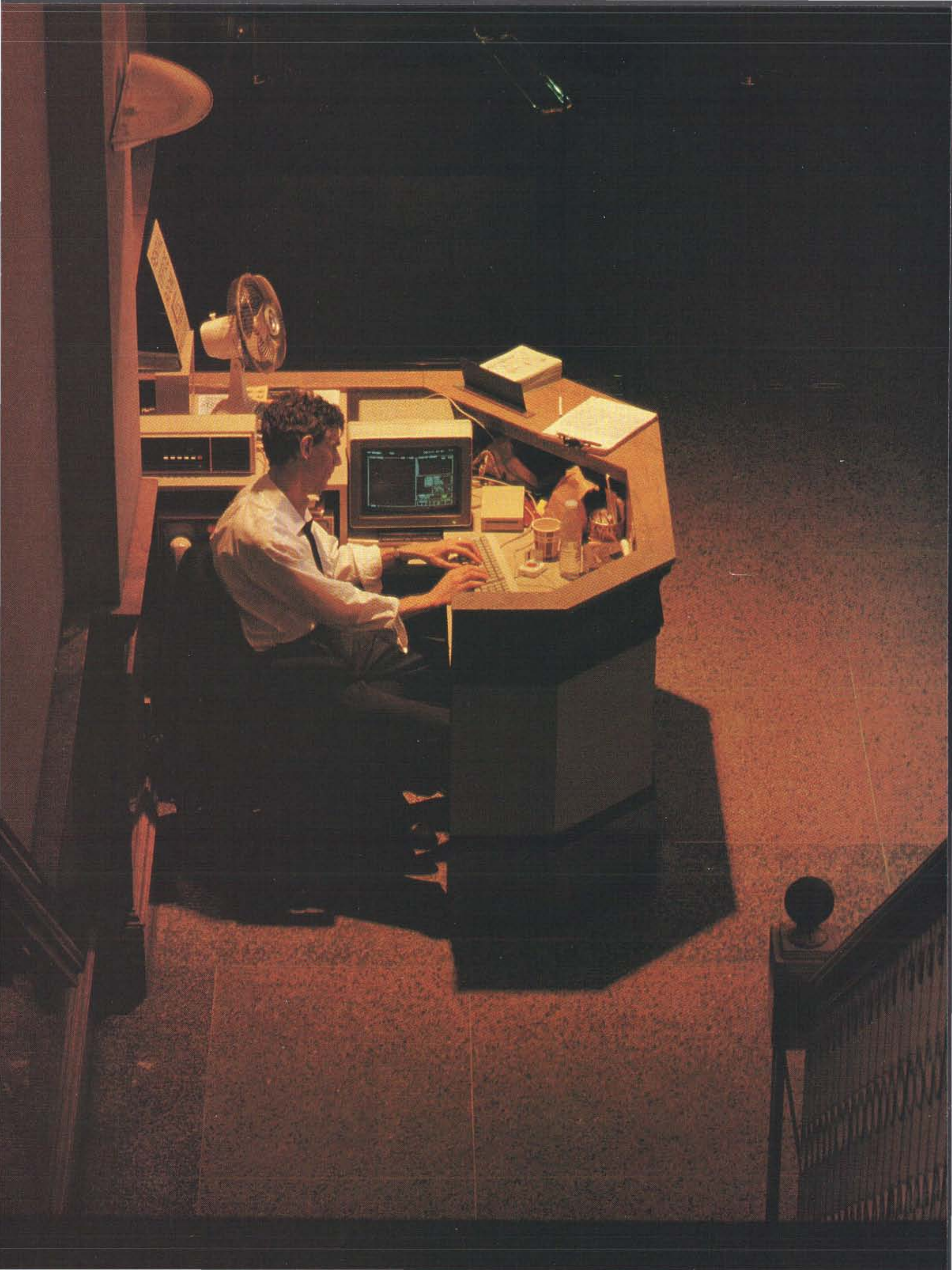
Our research relies on two approaches. Some questions can be studied through laboratory experiments. For instance, how do small groups respond emotionally to different forms of communication? Other questions, particularly those concerning organizational change, require field studies in actual organizations that have been routinely using computer networks. Data describing how hundreds of thousands of people currently use network communications can help predict how other people will work in the future as computer-based communications become more prevalent. Drawing on field studies and experiments, researchers gradually construct a body of evidence on how work and organizations are changing as network technology becomes more widely used. The process may sound straightforward, but in reality it is often full of exciting twists. People use technology in surprising ways, and effects often show up that contradict both theoretical predictions and managerial expectations.

One major surprise emerged as soon as the first large-scale computer network, known as the ARPANET, was begun in the late 1960s. The ARPANET was developed for the Advanced Research Projects Agency (ARPA), a part of the U.S. Department of Defense. ARPANET was intended to link computer scientists at universities and other research institutions to distant computers, thereby permitting efficient access to machines unavailable at the home institutions. A facility called electronic mail, which enabled researchers

LEE SPROULL and SARA KIESLER have spent more than a decade studying established electronic mail communities to learn about how they change patterns of communication within organizations. Sproull received a Ph.D. in social science from Stanford University in 1977. Until 1991, she was a professor of social sciences and sociology at Carnegie Mellon University. She is now a professor of management at Boston University's School of Management. Kiesler completed a Ph.D. in social psychology at Ohio State University in 1965. In 1979 she joined the faculty of Carnegie Mellon University, where she is a professor of social sciences and social psychology, as well as a member of the university's Human Computer Institute.

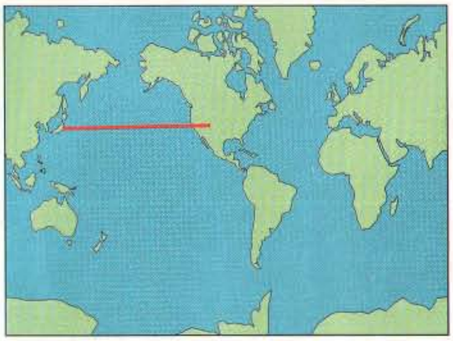
ISOLATED EMPLOYEES, such as a night watchman, benefit in particular from computer network links. Networks can create a web of social connections that stretch across time and that exist independently of an employee's physical location or hierarchical position. Because of the lack of social cues, people communicating electronically tend to talk more freely than they would in person. Networks potentially permit broader access to information and more democratic structures than are now found in most organizations. Exploiting that potential will force managers to grapple with issues of responsibility and control.

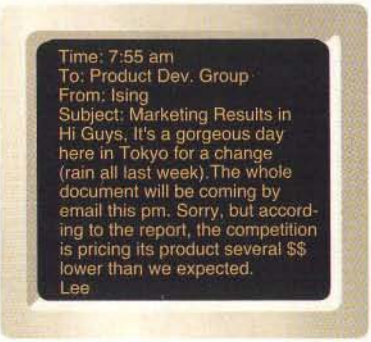
Reprinted from the September 1991 issue



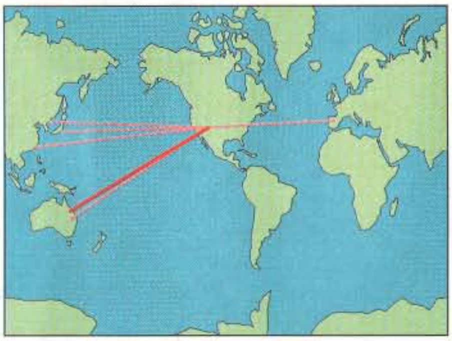
How People Work via Electronic Mail

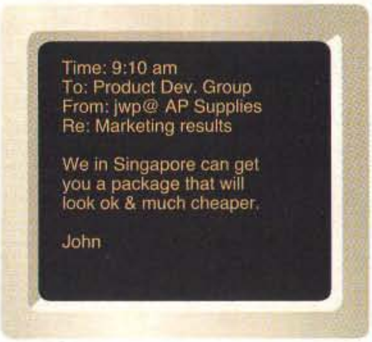
Global communication is possible using computer networks and electronic mail, as illustrated by the behavior of "Sue Jones," a composite based on a number of workers studied by the authors. On a typical day, Jones sends and receives from 25 to 100 pieces of electronic mail that cover both business and social topics. The maps record the links made during the course of the day, highlighting those associated with one of her working groups; the computer screens show the messages she sends to and receives from that working group. In this case, a potential crisis in the pricing of a new product is quickly resolved via electronic communication. Networks permit close, ongoing cooperation among co-workers who are physically located in very different corners of the globe.





Time: 7:55 am
To: Product Dev. Group
From: Ising
Subject: Marketing Results in
Hi Guys, It's a gorgeous day here in Tokyo for a change (rain all last week).The whole document will be coming by email this pm. Sorry, but according to the report, the competition is pricing its product several $$ lower than we expected.
Lee

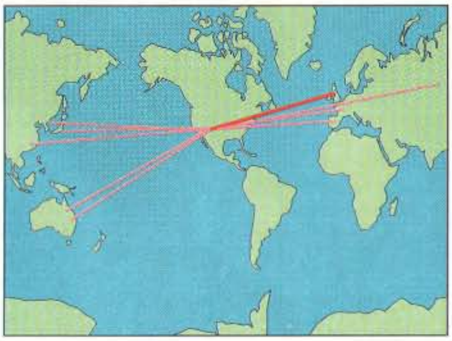


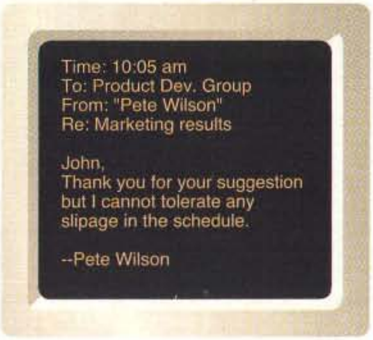


Time: 9:10 am
To: Product Dev. Group
From: jwp@ AP Supplies
Re: Marketing results

We in Singapore can get you a package that will look ok & much cheaper.

John

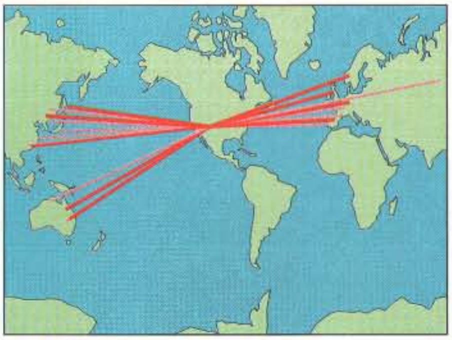


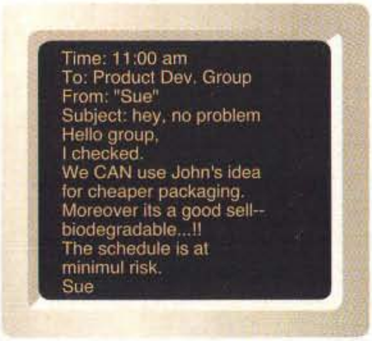


Time: 10:05 am
To: Product Dev. Group
From: "Pete Wilson"
Re: Marketing results

John,
Thank you for your suggestion but I cannot tolerate any slipage in the schedule.

--Pete Wilson





Time: 11:00 am
To: Product Dev. Group
From: "Sue"
Subject: hey, no problem
Hello group,
I checked.
We CAN use John's idea for cheaper packaging.
Moreover its a good sell--biodegradable...!!
The schedule is at minimul risk.
Sue

to communicate with one another, was considered a minor additional feature of the network.

Yet electronic mail rapidly became one of the most popular features of the ARPANET. Computer scientists around the country used ARPANET to exchange ideas spontaneously and casually. Graduate students discussed problems and shared skills with professors and other students without regard to their physical location. Heads of research projects used electronic mail to coordinate activities with project members and to stay in touch with other research teams and funding agencies. A network community quickly formed, filled with friends and collaborators who rarely, if ever, met in person. Although some administrators objected to electronic mail because they did not consider it a legitimate use of computer time, demand grew sharply for more and better network connections.

Since then, many organizations have adopted internal networks that link anywhere from a few to a few thousand employees. Some of these organizational networks have also been connected to the Internet, the successor to ARPANET. Electronic mail has continued to be one of the most popular features of these computer networks.

Anyone who has a computer account on a networked system can use electronic mail software to communicate with other users on the network. Electronic mail transmits messages to a recipient's electronic "mailbox." The sender can send a message simultaneously to several mailboxes by sending the message to a group name or to a distribution list. Electronic bulletin boards and electronic conferences are common variants of group electronic mail; they too have names to identify their topic or audience. Bulletin boards post messages in chronological order as they are received. Computer conferences arrange messages by topic and display grouped messages together.

The computer communications technology in most networked organizations today is fairly similar, but there exist large differences in people's actual communication behavior that stem from policy choices made by management. In some networked organizations, electronic mail access is easy and open. Most employees have networked terminals or computers on their desks, and anyone can send mail to anyone else. Electronic mail costs are considered part of general overhead expenses and are not charged to employees or to their departments. In the open-network organizations we have studied, people typically send and receive between 25 and 100 messages a day and belong to between 10 and 50 electronic groups. These figures hold across job categories, hierarchical position, age and even amount of computer experience.

In other networked organizations, managers have chosen to limit access or charge costs directly to users, leading to much lower usage rates. Paul Schreiber, a *Newsday* columnist, describes how his own organization changed from an open-access network to a limited-access one. Management apparently believed that reporters were spending too much time sending electronic mail; management therefore had the newspaper's electronic mail software modified so that reporters could still receive mail but could no longer send it. Editors, on the other hand, could still send electronic mail to everyone. Clearly, technology by itself does not impel change. Management choices and policies are equally influential.

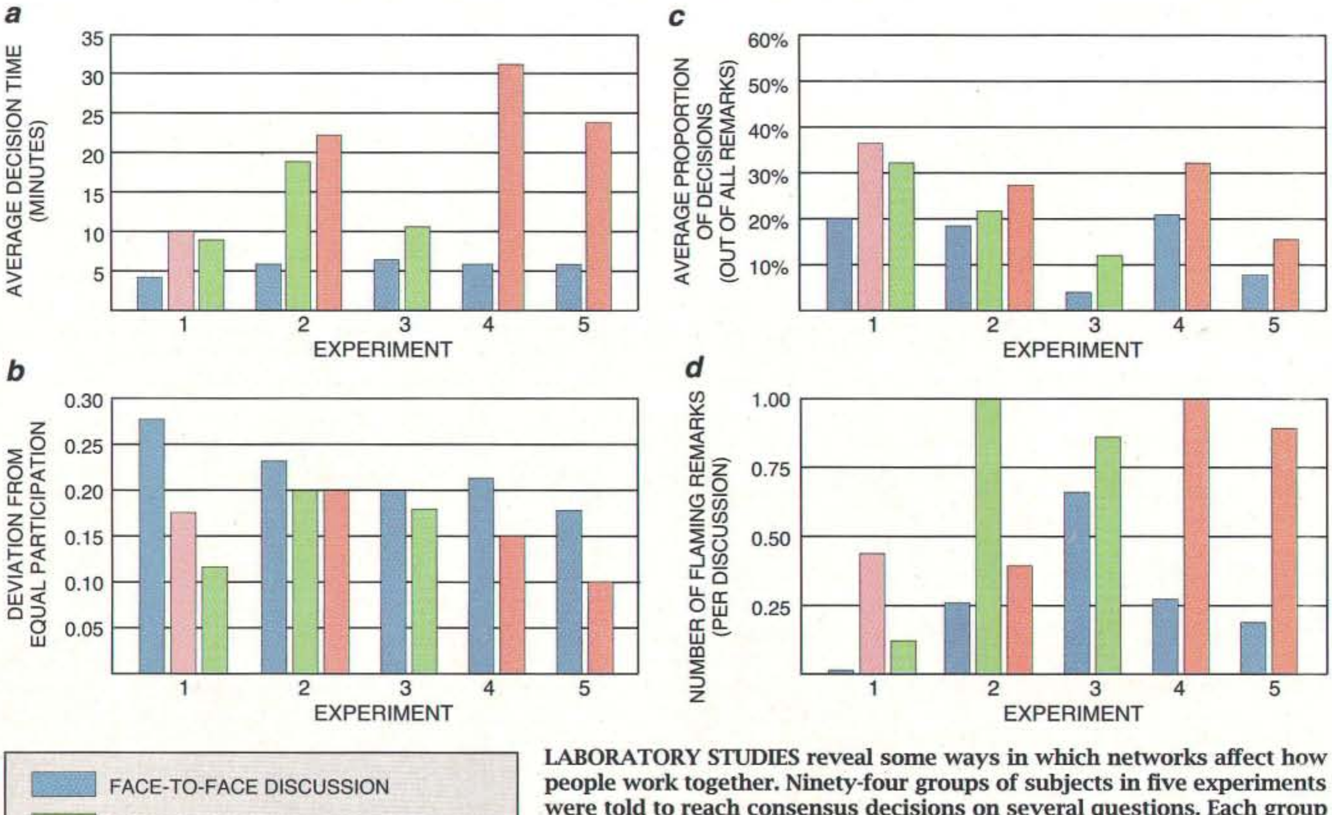
But even in organizations that have open access, anticipating the effect of networks on communication has proved no easy task. Some of the first researchers to study computer network communications thought the technology would improve group decision making over face-to-face discussion because computer messages were plain text. They reasoned that electronic discussions would be more purely intellectual, and so decision making would be less affected by people's social skills and personal idiosyncracies.

Research has revealed a more complicated picture. In an electronic exchange, the social and contextual cues that usually regulate and influence group dynamics are missing or attenuated. Electronic messages lack information regarding job titles, social importance, hierarchical position, race, age and appearance. The context also is poorly defined because formal and casual exchanges look essentially the same. People may have outside information about senders, receivers and situations, but few cues exist in the computer interaction itself to remind people of that knowledge.

In a series of experiments at Carnegie Mellon University, we compared how small groups make decisions using computer conferences, electronic mail and face-to-face discussion [*see illustration below*]. Using a network induced the participants to talk more frankly and more equally. Instead of one or two people doing most of the talking, as happens in many face-to-face groups, everyone had a more equal say. Furthermore, networked groups generated more proposals for action than did traditional ones.

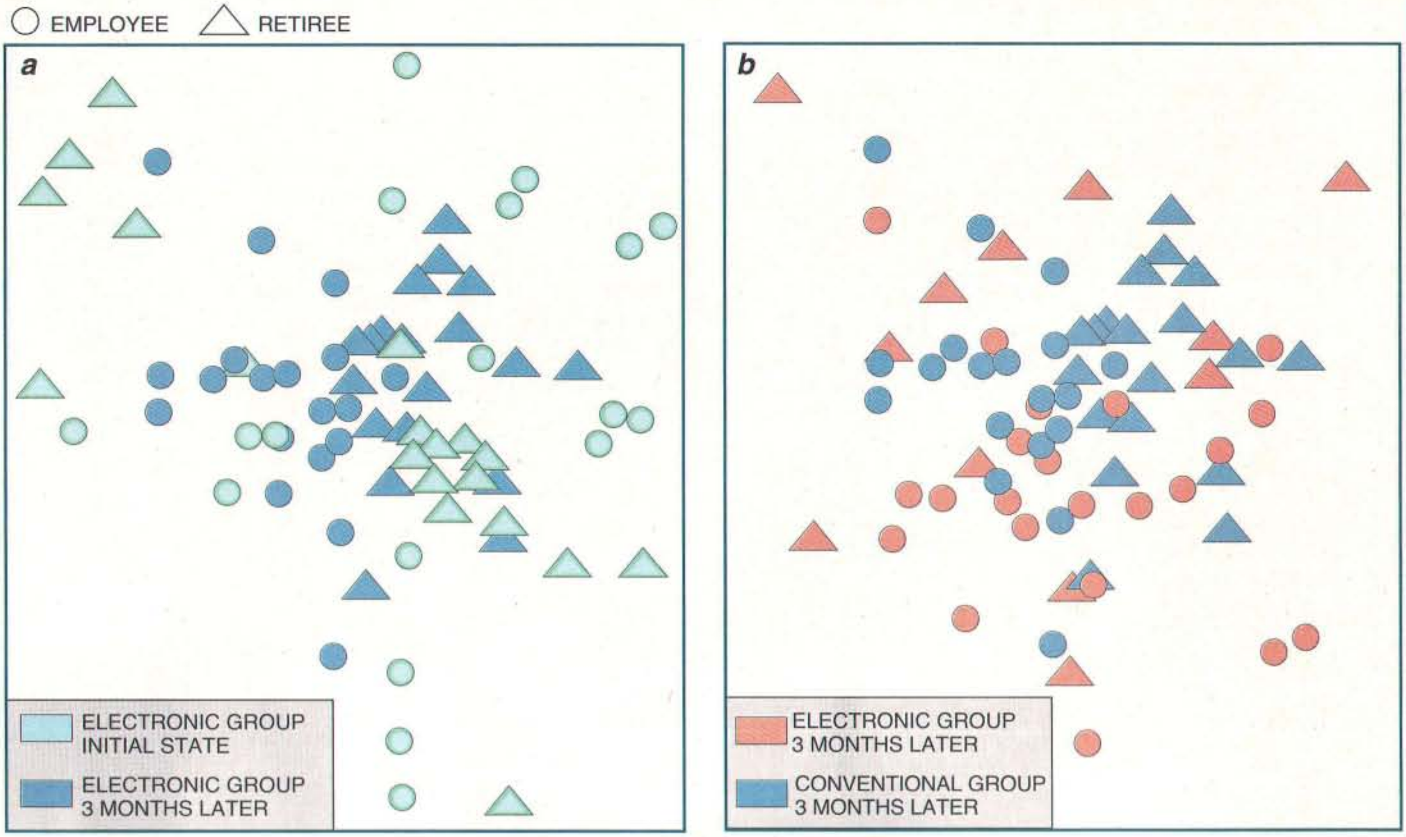
Open, free-ranging discourse has a dark side. The increased democracy associated with electronic interactions in our experiments interfered with decision making. We observed that three-person groups took approximately four times as long to reach a decision electronically as they did face-to-face. In one case, a group never succeeded in reaching consensus, and we were ultimately forced to terminate the experiment. Making it impossible for people to interrupt one another slowed decision making and increased conflict as a few members tried to dominate control of the network. We also found that people tended to express extreme opinions and vented anger more openly in an electronic face-off than when they sat together and talked. Computer scientists using the ARPANET have called this phenomenon "flaming."

We discovered that electronic communication can influence the effects of people's status. Social or job position normally is a powerful regulator of group interaction. Group members typ-



🟦 FACE-TO-FACE DISCUSSION
🟩 COMPUTER CONFERENCE (REAL TIME)
🟥 ELECTRONIC MAIL DISCUSSION
▨ ANONYMOUS ELECTRONIC MAIL

LABORATORY STUDIES reveal some ways in which networks affect how people work together. Ninety-four groups of subjects in five experiments were told to reach consensus decisions on several questions. Each group made some decisions electronically (through one or two network modes) and others face-to-face. When networked, all groups took longer to make a decision (*a*). On the other hand, they enjoyed more equal participation (*b*) and proposed more ideas (*c*). The electronic modes of discussion seemed to encourage "flaming," impassioned self-expression (*d*).



ically defer to those who have higher status and tend to follow their direction. Members' speech and demeanor become more formal in the presence of people who have high status. Higher-status people, in turn, talk more and influence group discussion more than do lower-status people.

Given that electronic conversations attenuate contextual cues, we expected that the effect of status differences within a group should also be reduced. In an experiment conducted with Vitaly Dubrovsky of Clarkson University and Beheruz Sethna of Lamar University, we asked groups containing high- and low-status members to make decisions both by electronic mail and face-to-face. The results confirmed that the proportion of talk and influence of higher-status people decreased when group members communicated by electronic mail.

Is this a good state of affairs? When higher-status members have less expertise, more democracy could improve decision making. If higher-status members truly are better qualified to make decisions, however, the results of consensus decisions may be less good.

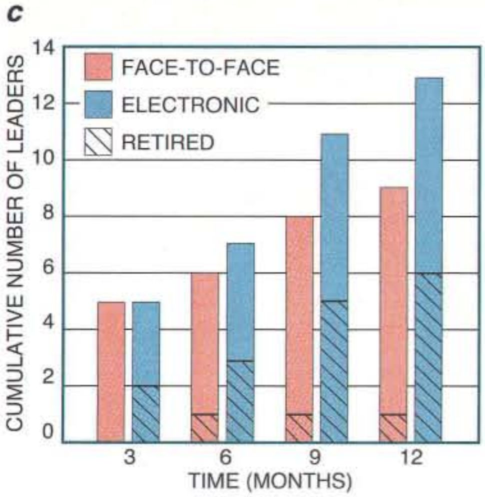
Shoshanah Zuboff of Harvard Business School documented reduced effects of status on a computer conference system in one firm. People who regarded themselves as physically unattractive reported feeling more lively and confident when they expressed themselves over the network. Others who had soft voices or small stature reported that they no longer had to struggle to be taken seriously in a meeting.

Researchers have advanced alternative explanations for the openness and democracy of electronic talk. One hypothesis is that people who like to use computers are childish or unruly, but this hypothesis does not explain experimental results showing that the same people talk more openly on a computer than when they are face-to-face. Another hypothesis holds that text messages require strong language to get a point across; this hypothesis explains flaming but not the reduction of social and status differences. The most promising explanation of the behavior of networked individuals is that when cues about social context are absent or weak, people ignore their social situation and cease to worry about how others evaluate them. Hence, they devote less time and effort to posturing and social niceties, and they may be more honest.

Researchers have demonstrated decreased social posturing in studies that ask people to describe their own behavior. In one of our experiments, people were asked to complete a self-evaluation questionnaire either by pencil and paper or via electronic mail. Those randomly assigned to reply electronically reported significantly more undesirable social behaviors, such as illegal drug use or petty crimes. John Greist and his colleagues at the University of Wisconsin found similar decreases in posturing when taking medical histories from clinical patients. People who responded to a computerized patient history interview revealed more socially and physically undesirable behavior than did those who answered the same questions asked by a physician.

These studies show that people are willing to reveal more about undesirable symptoms and behavior to a computer, but are these reports more truthful? An investigation of alcohol consumption conducted by Jennifer J. Waterton and John C. Duffy of the University of Edinburgh suggests an affirmative answer. In traditional surveys, people report drinking only about one half as much alcohol as alcohol sales figures would suggest. Waterton and Duffy compared computer interviews with personal interviews in a survey of alcohol consumption. People who were randomly assigned to answer the computer survey reported higher alcohol consumption than those who talked to the human interviewer. The computer-derived reports of consumption extrapolated more accurately to actual alcohol sales than did the face-to-face reports.

These and other controlled studies of electronic talk suggest that such communication is relatively impersonal, yet paradoxically, it can make people feel more comfortable about talk-



DYNAMIC GROUP STRUCTURES emerge when people converse electronically. These charts depict the behavior of two study groups, each containing a mix of employees and retired workers. One group worked in person, the other over a network. Members having the most information and social contacts appear near the center of the charts. Over time, the electronic group becomes more socially cohesive (*a*). Chart *b* compares the electronic group with the conventional group three months into the project and shows that retirees in particular become better integrated in the electronic group. Networks also encouraged more people to take on leadership roles (*c*).

ing. People are less shy and more playful in electronic discussions; they also express more opinions and ideas and vent more emotion.

Because of these behavioral effects, organizations are discovering applications for electronic group activities that nobody had anticipated. Computers can be valuable for counseling and conducting surveys about sensitive topics, situations in which many people are anxious and cover their true feelings and opinions. Networks are now being used for applications ranging from electronic Alcoholics Anonymous support groups to electronic quality circles.

Just as the dynamics of electronic communications differ from those conducted orally or by letters, so electronic groups are not just traditional groups whose members use computers. People in a networked organization are likely to belong to a number of electronic groups that span time zones and job categories. Some of these groups serve as extensions of existing work groups, providing a convenient way for members to communicate between face-to-face meetings. Other electronic groups gather together people who do not know one another personally and who may in fact have never had the opportunity to meet in person.

For example, Hewlett-Packard employs human-factors engineers who work in widely scattered locations around the world. These engineers may meet one another in person only once a year. An electronic conference creates ongoing meetings in which they can frequently and routinely discuss professional and company issues.

In some ways, electronic groups resemble nonelectronic social groups. They support sustained interactions, develop their own norms of behavior and generate peer pressure. Electronic groups often have more than 100 members, however, and involve relationships among people who do not know one another personally.

Employees whose organization is connected to the Internet or to a commercial network can belong to electronic groups whose members come from many different organizations. For example, Brian K. Reid of Digital Equipment Corporation reports that some 37,000 organizations are connected to USENET, a loosely organized network that exchanges more than 1,500 electronic discussion groups, called newsgroups. Reid estimates that 1.4 million people worldwide read at least one newsgroup.

Networked communication is only beginning to affect the structure of the workplace. The form of most current organizations has been dictated by the constraints of the nonelectronic world. Interdependent jobs must be situated in physical proximity. Formal command structures specify who reports to whom, who assigns tasks to whom and who has access to what information. These constraints reinforce the centralization of authority and shape the degree of information sharing, the number of organizational levels, the amount of interconnectivity and the structure of social relationships.

Organizations that incorporate computer networks could become more flexible and less hierarchical in structure. A field experiment conducted by Tora K. Bikson of Rand Corporation and John D. Eveland of Claremont Colleges supports the point. They formed two task forces in a large utility firm, each assigned to analyze employee retirement issues. Both groups contained 40 members, half of whom had recently retired from the company and half of whom were still employed but eligible for retirement. The only difference between the two groups was that one worked on networked computer facilities, whereas the other did not.

Both task forces created subcommittees, but the networked group created more of them and assigned people to more than one subcommittee. The networked group also organized its subcommittees in a complex, overlapping matrix structure. It added new subcommittees during the course of its work, and it decided to continue meeting even after its official one-year life span had ended. The networked task force also permitted greater input from the retirees, who were no longer located at the company. Although not every electronic group will be so flexible, eliminating the constraints of face-to-face meetings evidently facilitates trying out different forms of group organization.

Another effect of networking may be changed patterns of information sharing in organizations. Conventional organizations have formal systems of record keeping and of responsibilities for distributing information. Much of the information within an organization consists of personal experience that never appears in the formally authorized distribution system: the war stories told by service representatives (which do not appear in service manuals), the folklore about how the experimental apparatus really works (which does not appear in the journal articles) or the gossip about how workers should behave (which is not described in any personnel policy).

In the past, the spread of such personal information has been strongly determined by physical proximity and social acquaintance. As a result, distant or poorly connected employees have lacked access to local expertise; this untapped knowledge could represent an important informational resource in large organizations. Electronic groups provide a forum for sharing such expertise independent of spatial and social constraints.

One significant kind of information flow begins with the "Does anybody know...?" message that appears frequently on computer networks. A sender might broadcast an electronic request for information to an entire organization, to a particular distribution list or to a bulletin board. Anyone who sees the message can reply. We studied information inquiries on the network at Tandem Computers, Inc., in Cupertino, Calif., a computer company that employs 10,500 workers around the world. In a study we conducted with David Constant, we found an average of about six does-anybody-know messages broadcast every day to one company-wide distribution list.

Information requests typically come from field engineers or sales representatives who are soliciting personal experience or technical knowledge that they cannot find in formal documents or in their own workplace. At Tandem, about eight employees send electronic mail replies to the average question. Fewer than 15 percent of the people who answer a question are personally acquainted with the questioner or are even located in the same city.

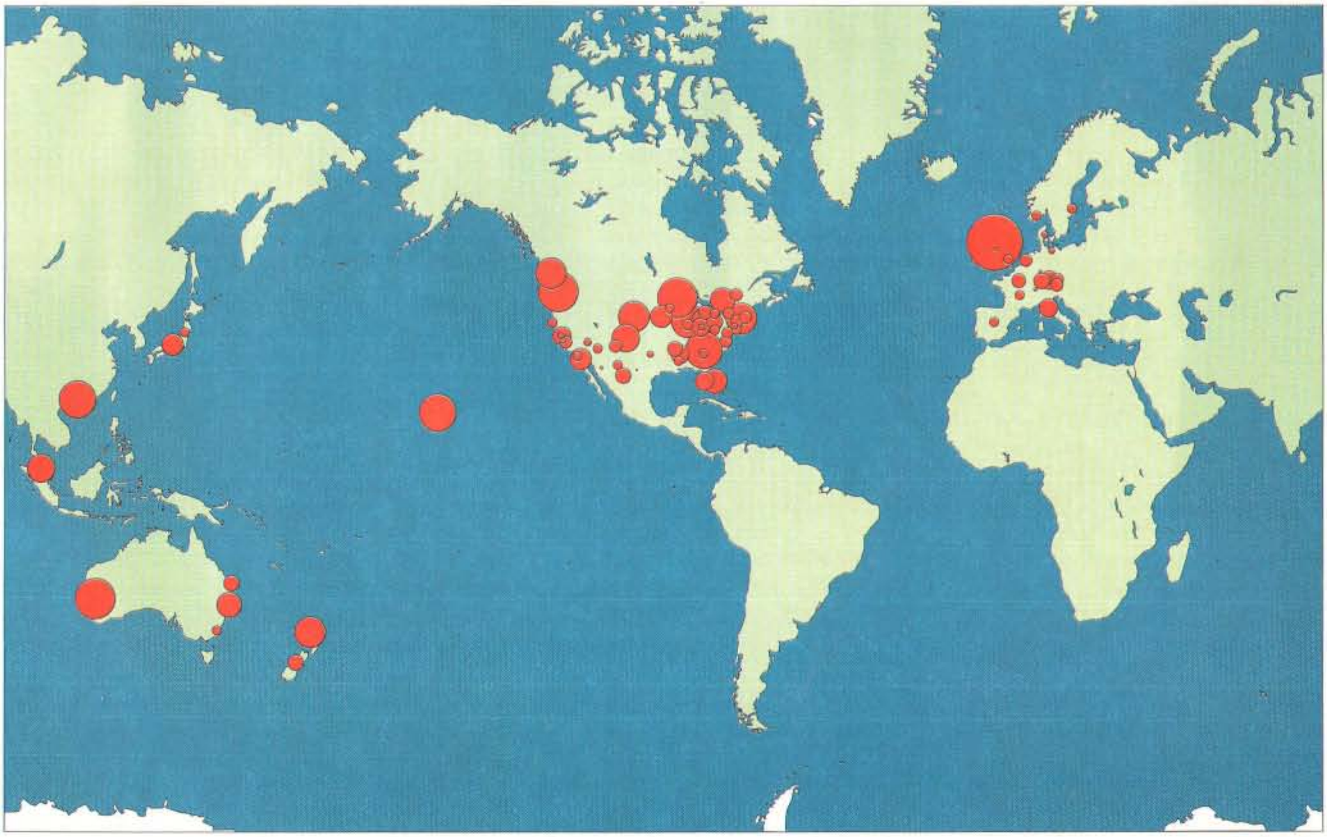
Question askers can electronically redistribute the answers they receive by putting them in a public computer file on the network. About half of the Tandem questioners make their reply files publicly available over the company network to other employees. Tandem takes this sharing process one step further by maintaining an electronic archive of question-and-reply files that is also accessible over the company network. The firm has thereby created a repository of information and working expertise that is endlessly accessible through space and time (for example, the expertise remains available when an employee is out of the office or he or she leaves the organization). A study by Thomas Finholt in our research program found that this archive is accessed more than 1,000 times a month by employees, especially those located in field offices away from the geographic center of the company.

The discretionary information sharing we discovered at Tandem and at other networked organizations seems to run contrary to nonelectronic behavior in organizations. The askers openly admit their ignorance to perhaps hundreds or even thousands of people. The repliers respond to requests for help from people they do not know with no expectation of any direct benefit to themselves.

One might wonder why people respond so readily to information requests made by strangers. Part of the explanation is that networks make the cost of responding extremely low in time and effort expended. Also, open-access networks favor the free flow of information. Respondents seem to believe that sharing information enhances the overall electronic community and leads to a richer information environment. The result is a kind of electronic altruism quite different from the fears that networks would weaken the social fabric of organizations.

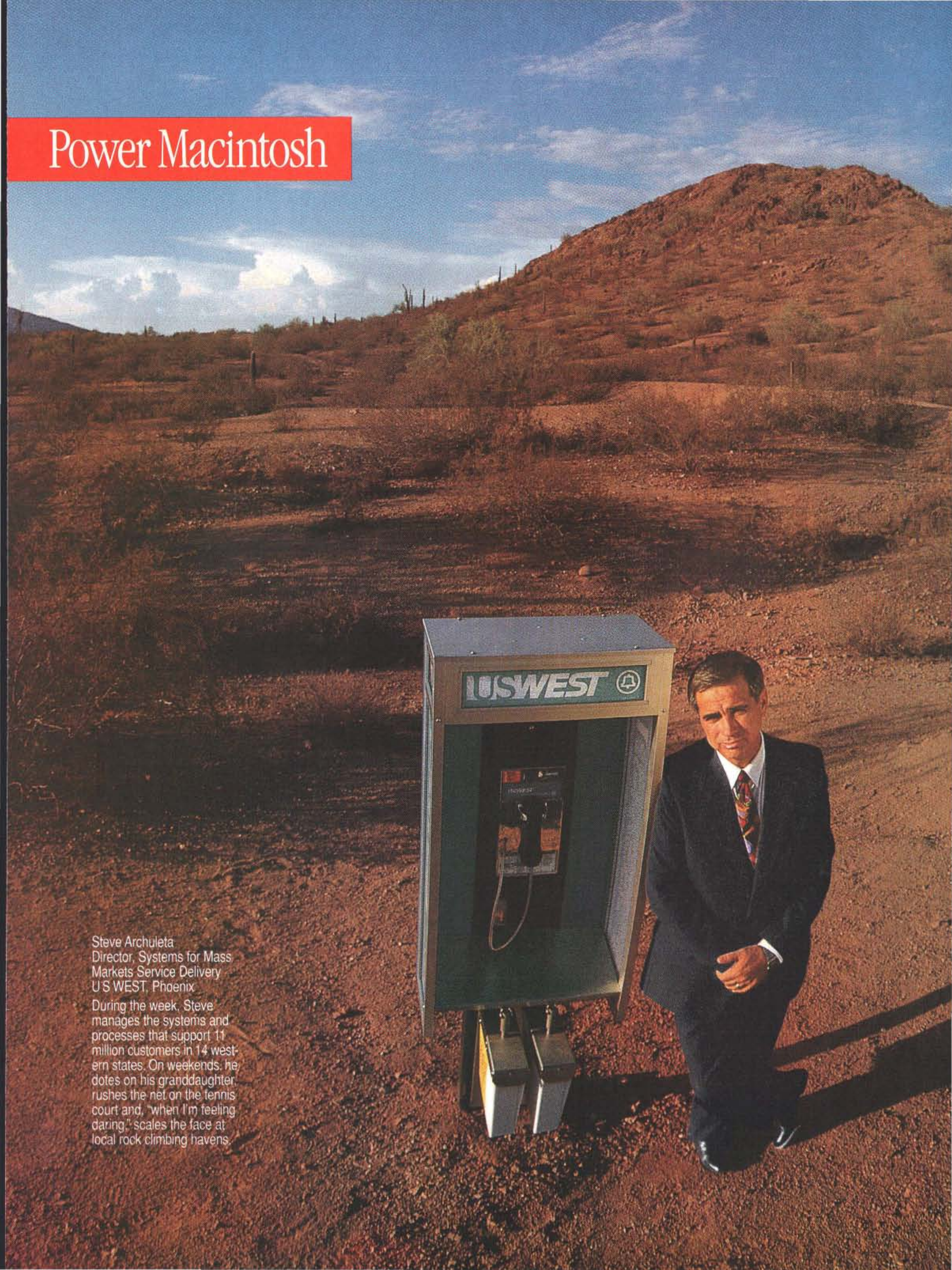
The changes in communication made possible by networks may substantially alter the relationship between an employee and his or her organization, the structure of organizations and the nature of management. Senior managers and key professionals usually have strong social and informational connections within their organizations and within their broader professional communities. Conversely, employees who reside on the organizational periphery by virtue of geographic location, job requirements or personal attributes have relatively few opportunities to make contact with other employees and colleagues.

Reducing the impediments to communication across both physical and social distance is likely to affect peripheral employees more than central ones. We, along with Charles Huff of St. Olaf College, studied this possibility for city employees in Fort Collins, Colo. Employees who used electronic mail extensively reported more commitment to their jobs and to their co-workers than did those who rarely used the network. This correlation was particularly strong for shift workers, who, because of the nature of their work, had fewer opportunities to see their colleagues than did regular day workers. As one policewoman told us,



ELECTRONIC LINKS have the greatest effect on workers located at outlying sites. Workers in field offices of Tandem Computers, Inc., whose headquarters are in Cupertino, Calif., have access to data files via a network. Circles indicate how many times each office tapped into one file (consisting of employee-initiated questions and answers about company products and services) over a one-year period; the greater the usage, the larger the circle. Workers in distant or isolated offices, where local expertise is relatively limited, made the most use of the information provided through the network.

“The teeth of one gear must mesh with those of another.”



Steve Archuleta
Director, Systems for Mass Markets Service Delivery
U S WEST, Phoenix

During the week, Steve manages the systems and processes that support 11 million customers in 14 western states. On weekends, he dotes on his granddaughter, rushes the net on the tennis court and, "when I'm feeling daring," scales the face at local rock climbing havens.

“There's leading edge. And bleeding edge. And you have to walk the line to stay competitive. That's why we use Power Macintosh. It's a RISC PC, so it's leading the way. But it's also a Mac that can run Windows and DOS, so it's compatible with everything we're doing now. That makes Power Macintosh a very safe choice. Which is good for a guy like me. I take my risks on the weekend.”

The RISC-based PowerPC™ microprocessor at the heart of every Power Macintosh takes desktop computing to new levels of performance and compatibility. For the name of your nearest authorized Apple reseller, call 800-732-3131, ext. 750, today.

"In some cases, we used to have three different PCs on one desk–an absolute monument to systems failure. So that's where we put the Power Macintosh first. Now there's one PC per desk, the Power Macintosh. One happy user per desk, too."

Mac™OS For Windows

Power Macintosh™ fits right in at U S WEST, where people use many different types of PCs and 30 or so custom DOS and Windows programs. "Power Macintosh helps our people collaborate, no matter what kind of PC they use."

Power Macintosh™. The business Macintosh™.

Apple

Power Macintosh. Compatible today. Compatible tomorrow.

Any computer can claim it's compatible—with the software written specifically for it, with the hardware designed to coexist with its particular bus architecture, and so on.

Most PCs can also claim compatibility with other PCs designed to the same standards.

But only one computer can claim true cross-platform compatibility right out of the box.* Compatibility that crosses technological barriers—and cultural chasms. Compatibility that serves your most immediate needs. While keeping options open for the future.

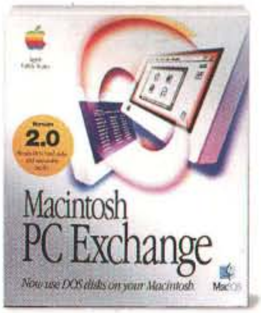
Which computer is that? Power Macintosh, of course.

Introducing competitive coexistence.

It's no revelation to anyone that Macintosh® computers exist alongside many other kinds of computers. Indeed, more people than not have chosen to use computers that run DOS and Windows.

And, while we'll stand fast in our belief that Macintosh is the most intuitive and productive way to get anything done, we know that everyone needs the ability to work in a mixed environment.

That's why every Power Macintosh reads, writes and formats disks for DOS, Windows and OS/2 (in addition to disks for Macintosh, of course). And why every Power Macintosh also includes Macintosh PC Exchange.™ So you can open files created under DOS and Windows—even if you don't have the applications that created them.



That's also why you should check out Insignia Solutions' SoftWindows™—the software that actually lets you run Windows and DOS on your Macintosh. Which means, of course, your Mac™ can run all kinds of programs designed for DOS and Windows, too.

And if you run UNIX®? Then you'll appreciate the UNIX compatibility provided by our MacX™ extension to the Mac OS.

We've got it wired.

But we don't stop there. For bona fide DOS and Windows fans, or for Mac fans who find it necessary to use DOS and Windows on a regular basis, we also build a Power Macintosh that has a 486 DX2/66 built in. (We call it the Power Macintosh 6100/66 DOS Compatible for short.) So with one touch of a key, you can switch from the Mac OS to Windows. Or DOS. And back.

And for those of you who have already bought a Power Mac™ 6100, adding this level of hardware compatibility will soon be as easy as plugging a card into any available slot.

A future powered by PowerPC.

What makes all this possible? The innovative new microprocessor at the heart of Power Macintosh: the RISC-based PowerPC.

Is PowerPC just another chip? No. It's the result of a far-reaching collaboration between Apple, IBM and Motorola.

It successfully brings the power of Reduced Instruction Set Computing to an affordable desktop computer. It runs cool. Efficiently. Flawlessly. And fast. So fast, in fact, that popular business, scientific and graphics programs optimized for PowerPC run two to eight times faster than they ran on any previous Mac.

And that's just today. Future generations of PowerPC will be even faster. Run even cooler. Fuel even more powerful applications. Build new bridges of compatibility—between computer platforms. And between the people who use them.

Which is something not every microprocessor can claim. Because while even the newest microprocessors based on CISC technology are quickly reaching a performance plateau, PowerPC is just taking off.

IBM®

MOTOROLA

To get more information by fax about this and other critical compatibility issues, call us at 800-770-4852. Or, if you prefer, contact us via the Internet at http://www.info.apple.com.

Apple.

*The Power Macintosh 6100/66 DOS Compatible runs DOS, Windows and Mac OS. Other Power Macintosh models require SoftWindows by Insignia Solutions to run DOS and Windows systems, and SoftWindows may need to be purchased separately. © 1995 Apple Computer, Inc. All rights reserved. Apple, the Apple logo and Macintosh are registered trademarks of Apple Computer, Inc. Mac, the Mac OS logo, Macintosh PC Exchange, MacX, Power Mac and Power Macintosh are trademarks of Apple Computer, Inc. PowerPC is a trademark of International Business Machines Corporation, used under license therefrom. Windows is a trademark of Microsoft Corporation and SoftWindows is a trademark used under license by Insignia from Microsoft Corporation. All other product names and logos are trademarks or registered trademarks of their respective companies.

"Working the night shift, it used to be that I would hear about promotions after they happened, though I had a right to be included in the discussion. Now I have a say in the decision making."

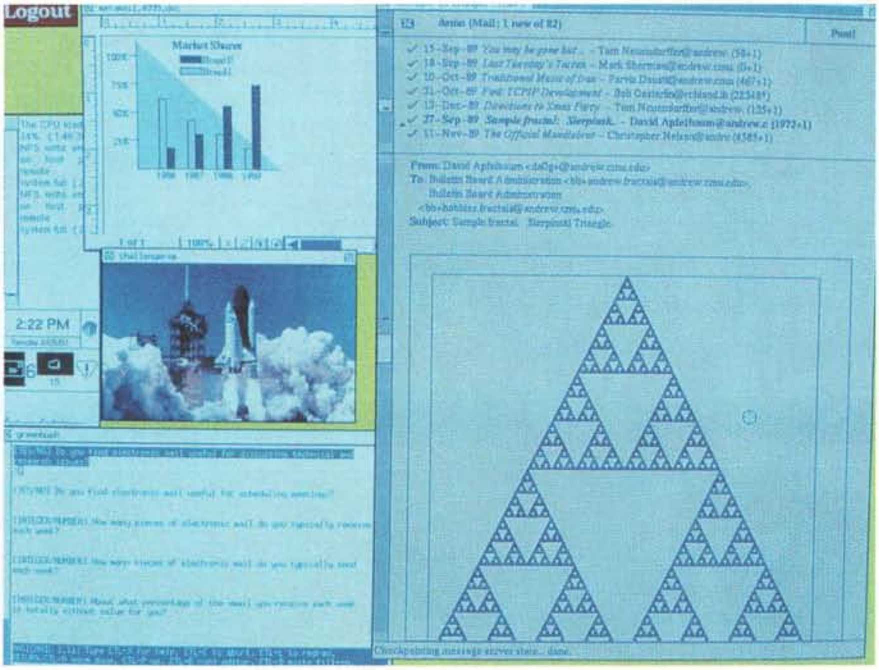
Organizations are traditionally built around two key concepts: hierarchical decomposition of goals and tasks and the stability of employee relationships over time. In the fully networked organization that may become increasingly common in the future, task structures may be much more flexible and dynamic. Hierarchy will not vanish, but it will be augmented by distributed lattices of interconnections.

In today's organizations, executives generally know whom they manage and manage whom they know. In the future, however, managers of some electronic project groups will face the challenge of working with people they have never met. Allocating resources to projects and assigning credit and blame for performance will become more complex. People will often belong to many different groups and will be able to reach out across the network to acquire resources without management intervention or perhaps even without management knowledge.

A recent case in mathematics research hints at the nature of what may lie ahead. Mathematicians at Bell Communications Research (Bellcore) and at Digital Equipment sought to factor a large, theoretically interesting number known as the 9th Fermat number. They broadcast a message on the Internet to recruit researchers from universities, government laboratories and corporations to assist them in their project. The several hundred researchers who volunteered to help received—via electronic mail—software and a piece of the problem to solve; they also returned their solutions through electronic mail.

After results from all the volunteers were combined, the message announcing the final results of the project contained a charming admission:

> We'd like to thank everyone who contributed computing cycles to this project, but I can't: we only have records of the person at each site who installed and managed the code. If you helped us, we'd be delighted to hear from you; please send us your name as you would like it to appear in the final version of the paper. (Broadcast message from Mark S. Manasse, June 15, 1990.)



ADVANCED ELECTRONIC MAIL, such as this experimental setup at Bellcore, will make it possible to transmit graphics, animation, self-contained computer programs and audio recordings. Such features will broaden the applications for networked communications. They may also restore some of the social cues now absent from electronic mail and thereby modify the behavior of its users.

Networking in most organizations today is limited to data communications, often for economic or financial applications such as electronic data interchange, electronic funds transfer or remote transaction processing. Most organizations have not yet begun to confront the opportunities and challenges afforded by connecting their employees through networks.

Among those that have, managers have responded in a variety of ways to changes that affect their authority and control. Some managers have installed networks for efficiency reasons but ignored their potential for more profound changes. Some have restricted who can send mail or have shut down electronic discussion groups. Others have encouraged using the network for broadening participation and involving more people in the decision-making process. The last actions push responsibility down and through the organization and also produce their own managerial issues.

A democratic organization requires competent, committed, responsible employees. It requires new ways of allocating credit. It increases unpredictability, both for creative ideas and for inappropriate behavior. Managers will have to come up with new kinds of worker incentives and organizational structures to handle these changes.

The technology of networks is changing rapidly. Electronic mail that includes graphics, pictures, sound and video will eventually become widely available. These advances will make it possible to reintroduce some of the social context cues absent in current electronic communications. Even so, electronic interactions will never duplicate those conducted face-to-face.

As more people have ready access to network communications, the number and size of electronic groups will expand dramatically. It is up to management to make and shape connections. The organization of the future will depend significantly not just on how the technology of networking evolves but also on how managers seize the opportunity it presents for transforming the structure of work.

FURTHER READING

WHAT DO COMPUTERS DO? James Rule and Paul Attewell in *Social Problems*, Vol. 36, No. 3, pages 225–241; June 1989.

INTELLECTUAL TEAMWORK: THE SOCIAL AND TECHNOLOGICAL FOUNDATIONS OF COOPERATIVE WORK. Edited by Jolene Galegher, Robert E. Kraut and Carmen Egido. Lawrence Erlbaum Associates, 1990.

CONNECTIONS: NEW WAYS OF WORKING IN THE NETWORKED ORGANIZATION. Lee Sproull and Sara Kiesler. MIT Press, 1991.

SOCIAL RELATIONSHIPS IN ELECTRONIC COMMUNITIES. Charles Dunlop and Rob Kling in *Computerization and Controversy: Value Conflicts and Social Choices.* Edited by Charles Dunlop and Rob Kling. Academic Press, 1991.

Computers, Networks and the Corporation

Computer networks are forging new kinds of markets and new ways to manage organizations. The result will be a major change in corporate structure and management style

by Thomas W. Malone and John F. Rockart

About 150 years ago the economy in the U.S. and Europe began to undergo a period of change more profound than any experienced since the end of the Middle Ages. We call that change the Industrial Revolution. The industrial economies are now in the early stages of another transformation that may ultimately be at least as significant.

There is a critical difference this time, however. Changes in the economies of production and transportation drove the revolution of the last century. The revolution under way today will be driven not by changes in production but by changes in coordination. Whenever people work together, they must somehow communicate, make decisions, allocate resources and get products and services to the right place at the right time. Managers, clerks, salespeople, buyers, brokers, accountants—in fact, almost everyone who works—must perform coordination activities.

It is in these heavily information-based activities that information technologies have some of their most important uses, and it is here that they will have their most profound effects. By dramatically reducing the costs of coordination and increasing its speed and quality, these new technologies will enable people to coordinate more effectively, to do much more coordination and to form new, coordination-intensive business structures.

The core of the new technologies is the networked computer. The very name "computer" suggests how one usually thinks of the device—as a machine for computing, that is, for taking in information, performing calculations and then presenting the results. But this image of computing does not capture the essence of how computers are used now and how they will be used even more in the future. Many of the most important uses of computers today are for coordination tasks, such as keeping track of orders, inventory and accounts. Furthermore, as computers become increasingly connected to one another, people will find many more ways to coordinate their work. In short, computers and computer networks may well be remembered not as technology used primarily to compute but as coordination technology.

To understand what is likely to happen as information technology improves and its costs decline, consider an analogy with a different technology: transportation. A first-order effect of transportation technology was simply the substitution of new transportation technologies for the old. People began to ride in trains and automobiles rather than on horses and in horse-drawn carriages.

As transportation technology continued to improve, people did not use it just to substitute for the transportation they had been using all along. Instead a second-order effect emerged: people began to travel more. They commuted farther to work each day. They were more likely to attend distant business meetings and to visit faraway friends and relatives.

Then, as people used more and more transportation, a third-order effect eventually occurred: the emergence of new "transportation-intensive" social and economic structures. These structures, such as suburbs and shopping malls, would not have been possible without the wide availability of cheap and convenient transportation.

Improved coordination technology has analogous effects. A first-order effect of reducing coordination costs is the substitution of information technology for human coordination. For example, data-processing systems helped to eliminate thousands of clerks from the back offices of insurance companies and banks. Similarly, computer-based systems have replaced scores of factory "expediters." Today computers track the priority of each job in the factory and indicate the most critical ones at each workstation. More generally, the long-standing prediction that computers will lead to the demise of middle management finally seems to be coming true. In the 1980s many companies flattened their managerial hierarchies by eliminating layers of middle managers.

A second-order effect of reducing coordination costs is an increase in the overall amount of coordination used. For instance, contemporary airline res-

THOMAS W. MALONE and JOHN F. ROCKART, both at the Massachusetts Institute of Technology, study the effects of computer technology on business. Malone, who directs the Center for Coordination Science at M.I.T., received his Ph.D. from Stanford University and holds degrees in applied mathematics, engineering and psychology. He was previously a research scientist at the Xerox Palo Alto Research Center. Rockart received his M.B.A. from Harvard University and his Ph.D. from M.I.T., where he directs the Center for Information Systems Research. He serves on the board of directors of five organizations and lectures at several major companies.

TRADING FLOOR of the Paris Stock Exchange bustles with activity (*top*), much as do those of other exchanges. In contrast, the floor of the London International Stock Exchange, where trading is electronic, is quite leisured in pace (*bottom*). The distinction reflects how networks and coordination technology have begun to restructure the way businesses conduct transactions.

Reprinted from the September 1991 issue

Groupe D
Groupe E

13 08

ervation systems enable travel agents to consider more flight possibilities for a given customer more easily. These systems have led to an explosion of special fares and price adjustments. American Airlines and United Air Lines, which provide the largest systems, have benefited significantly from the fees they charge for this service. For instance, in 1988 American made about $134 million from its reservation system—almost 15 percent of its total income. In addition, access to up-to-the-minute information about ticket sales on all airlines enables American and United to adjust their fare schedules to maximize profits.

Otis Elevator Company also increased the amount of its coordination—primarily to improve maintenance service for its customers. With its Otisline system, highly trained multilingual operators receive trouble calls through a national toll-free number. The operators record the problems in a computer data base and then electronically dispatch local repair people.

This real-time availability of data has vastly improved the management of repair activities. For instance, if a particular type of part has failed during the past week on eight of 100 elevators, Otis can preemptively replace that part on the other 92 elevators. Although this kind of nationwide correlation of data was possible before, the degree of communication and coordination required was impractical. These capabilities have played a major role in reducing maintenance calls by nearly 20 percent.

In some instances, the second-order effect of an increase in demand may overwhelm the first-order effect of substitution. For example, in one case we studied, a computer conferencing system helped to remove a layer of middle managers. Several years later, however, almost the same number of new positions (for different people at the same grade level) had been created. According to people in the company, the new specialists took on projects not considered before. Evidently, managerial resources no longer needed for simple communication could now be focused on more complex coordination tasks.

A third-order effect of reducing coordination costs is a shift toward the use of more coordination-intensive structures. A prime example is Frito-Lay, Inc., studied by Lynda M. Applegate of Harvard Business School and others. At Frito-Lay, some 10,000 route salespeople record all sales of each of 200 grocery products on hand-held computers as they deliver goods to customers on their route. Each night, the stored information is transmitted to a central computer. In turn, the central computer sends information on changes in pricing and product promotions to the hand-held computers for use the next day. Each week, the main computer summarizes the centrally stored information and combines it with external data about the sales of competitive brands. Some 40 senior executives and others can then gain access to this information through an executive information system (EIS).

The availability of these data has enabled Frito-Lay to push key decisions down from corporate headquarters to four area heads and several dozen district managers. The managers can use the data not only to compare actual sales to sales targets but also to recommend changes in sales strategy to top management. This entire coordination-intensive structure has become possible only in the past few years because of the improved capability and reduced costs of hand-held computers, EIS software, computer cycles and telecommunications equipment.

Coordination-intensive structures do not just link different people in the

RESERVATION CENTER of Rosenbluth Travel in Philadelphia transmits travel information to a "back room," which coordinates activities of other agencies from around the world. Information about demand becomes available immediately, enabling the connected firms to respond to the market more effectively and profitably than can isolated agencies.

same companies. Many of the most interesting new structures involve links among different companies. For example, the U.S. textile industry has begun implementing a series of electronic connections among companies as part of the Quick Response program. As described by Janice H. Hammond of Harvard Business School and others, these electronic connections link companies all along the production chain, from suppliers of fibers (such as wool and cotton) to the mills that weave these fibers into fabric, to the factories that sew garments and, ultimately, to the stores that sell the garments to consumers.

When such networks are fully implemented, they will help companies respond quickly to demand. For instance, when a sweater is sold in New York City, a scanner reading the bar-coded label may automatically trigger ordering, shipping and production activities all the way back to the wool warehouse in South Carolina. This new, multiorganizational structure will reduce inventory costs throughout the value chain. The textile-apparel retail industry spends about $25 billion in inventory costs every year; the Quick Response approach may save half that amount.

Wal-Mart has already established parts of a similar system that links the retailer to Procter & Gamble Company and several of its other major suppliers. In doing so, Wal-Mart has eliminated significant parts of its own purchasing groups and contracted with its suppliers to replace products as they are sold. In one such experiment, both unit sales and inventory turnover increased by about 30 percent.

Sometimes technology helps to create interorganizational networks—not just among buyers and suppliers but also among potential competitors. For example, Eric K. Clemons of the University of Pennsylvania has studied the Rosenbluth International Alliance, a consortium of travel agencies around the world that share customer records, services and software. The alliance also provides clients with toll-free English-language help lines in every major country. This consortium of independent agencies, led by Rosenbluth Travel in Philadelphia, can therefore manage all travel arrangements for international trips and for meetings of people from many parts of the globe.

The textile firms near Prato, Italy, illustrate a related kind of interoganizational alliance. As described by Michael J. Piore and Charles F. Sabel of the Massachusetts Institute of Technology, the operation of a few large textile mills was broken into many small firms, coordinated in part by electronic connections among them. This network can flexibly adjust to changes in demand, sometimes shifting orders from an overloaded mill to one with spare capacity. The structure also takes advantage of the entrepreneurial motivation of the owners: in small mills, the owners' rewards are more closely linked to their own efforts than is the case in large ones.

ROUTE SALESMAN of Frito-Lay, Inc., enters inventory data into a hand-held computer. The information will be combined with similar data from 10,000 other salespeople and made accessible to management the next morning. The rapid availability of data has enabled Frito-Lay to decentralize pricing and inventory decisions. Instead of corporate headquarters, four area heads and the district managers set prices, and the salespeople determine the product mix.

As these examples show, information technology is already facilitating the emergence of new, coordination-intensive structures. What do these changes mean for the organizations of the near future?

A surprising result of our research is a prediction that information technology should lead to an overall shift from internal decisions within firms toward the use of markets to coordinate economic activity. To see why, consider that all organizations must choose between making the goods or services they need and buying them from outside suppliers. For instance, General Motors Corporation must decide whether to make tires internally or purchase them from a tire manufacturer.

Each of these two forms of coordination—internal and external—has advantages and disadvantages. As Oliver Williamson of the University of California at Berkeley and others have argued, buying things from an outside supplier often requires more coordination than making them internally. To buy tires, General Motors may need to compare many potential suppliers, negotiate contracts and do formal accounting for the money that changes hands. Coordinating the production of tires internally, on the other hand, can often be done less formally and at lower cost, with telephone calls and meetings.

But improved information technology should reduce the costs of both internal and external coordination, much as transportation technology lowered the expense of traveling. When trains and automobiles reduced the difficulty of traveling, more people chose to live in the suburbs rather than in the cities to reap such benefits as extra living space. Similarly, when information technology reduces the costs of a given amount of coordination, companies will choose to buy more and make less. The additional coordination required in buying will no longer be as expensive, and buying has certain advantages. For instance, when General Motors buys tires, it can take advantage of the supplier's economies of scale and pick the best tires currently available from any supplier whenever its needs change. Thus, we expect networks to lead to less vertical integration—more buying

SCANNING OF BAR CODES is one component of the textile industry's Quick Response program. The scans provide information about the contents of a package to each member of the distribution chain, from supplier to retailer. The rapid transmission of data reduces inventory and enables the firms to adjust quickly to demand.

rather than making—and to the proliferation of smaller firms. More electronically mediated alliances (such as the Rosenbluth International Alliance) and an increased use of electronic markets to pick suppliers (such as the airline reservation system) will result.

This argument implies that information technology will help make markets more efficient. Buyers will no longer have to exert great effort to compare products and prices from many different suppliers. Instead an electronic market can easily and inexpensively collect and distribute such information.

These more efficient markets threaten firms whose strategic advantages rest on market inefficiencies. For instance, as Clemons described, when the London International Stock Exchange installed an electronic trading system, the trading floor became virtually deserted within weeks. Trading moved to electronic terminals around the world. The system greatly reduced the costs of matching buyers and sellers. This change, in turn, dramatically reduced the profits of brokers and trading specialists, who previously had had a monopoly on performing this function. The potential decline in profit may explain why many other exchanges still resist electronic trading.

Many other kinds of intermediaries, such as distributors and retailers, are becoming vulnerable as well. For example, consumers can now bypass retail stores entirely by using computer-based systems such as Comp-U-Card and Comp-u-store to buy household goods and services at substantial savings. Electronic markets can also make evaluating product quality easier; we expect that it is only a matter of time before networks contain extensive comments and evaluations from previous buyers, becoming a kind of instantaneous, on-line *Consumer Reports.*

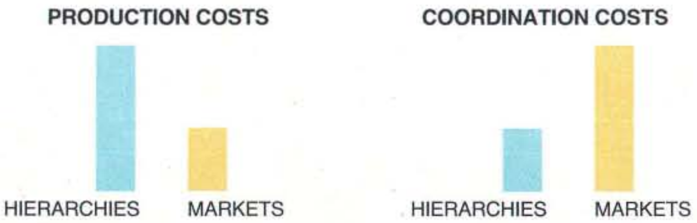
Increasing market efficiency also implies that firms should focus more carefully on the few core competencies that give them strategic advantages in the marketplace. They should buy the additional, more peripheral products and services they need instead of making them. For instance, in the past few years, Ford Motor Company and Chrysler Corporation have significantly increased their proportion of externally purchased components, such as tires and batteries.

Even though information technology can be strategically important, single innovations in information technology are seldom in themselves a source of continuing competitive advantage. For example, American Hospital Supply (now Baxter Healthcare Corporation) won high praise for its early system that let customers place orders electronically without requiring a salesperson. This system made ordering from American Hospital easier than doing so from competitors and reduced the time salespeople had to spend on the clerical aspects of taking an order. But contrary to original expectations, systems like these do not "lock in" customers in the long run. Instead customers eventually seem to prefer electronic systems that provide a choice among several vendors. Similarly, an automatic teller machine system that may once have been a competitive advantage for a bank is now largely a competitive necessity.

One way to maintain an upper hand is to keep innovating so rapidly that other firms always lag a step behind. Another way, as Clemons has noted, is to use information technology to leverage some other structural advantage. For instance, Barclay deZoete Wedd, a British brokerage firm, continues to benefit from an electronic stock-trading system because the company already controlled the trading of far more stocks than did any of their competitors.

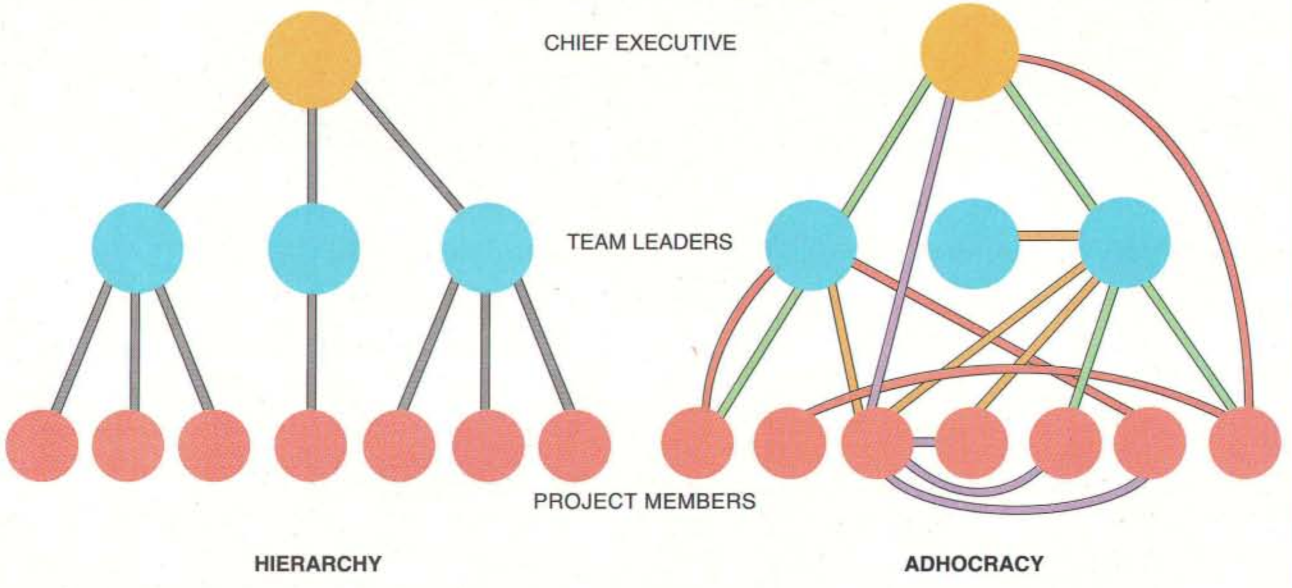
In addition to markets, another coordination-intensive organizational structure likely to become much more com-

Relative Costs for Hierarchies and Markets



Making products in vertically integrated hierarchies usually involves higher production costs than buying them in the market. Buying from an outside supplier allows a company to exploit economies of scale and other production cost advantages. But buying usually requires higher coordination costs; a firm must find a supplier, negotiate contracts and account for payments.

Corporate Structures



mon is what some management theorists call a networked organization, or, more picturesquely, an "adhocracy," a term Alvin Toffler popularized in his book *Future Shock*. This form is already common in organizations such as law firms, consulting companies and research universities. Such organizations and institutions must continually readjust to a changing array of projects, each requiring a somewhat different combination of skills and other resources. These organizations depend on many rapidly shifting project teams and much lateral communication among these relatively autonomous, entrepreneurial groups.

The adhocracy contrasts with the conventional business organization of today: the hierarchy. Hierarchies are common partly because they provide a very economical way of coordinating large numbers of people. In principle, decision makers in a hierarchy can consider all the information known to anyone in the group with much less communication than would be needed if each person communicated with everyone else.

In practice, however, hierarchies have severe limitations. Central decision makers can become overloaded and therefore unable to cope effectively with rapidly changing environments or to consider enough information about complex issues. Furthermore, people at the bottom may feel left out of the decision making and as a result be less motivated to contribute their efforts.

As information technology reduces communication costs, the nonhierarchical structures (such as markets and adhocracies) may help overcome the limitations of hierarchies. For example, because of the large amount of unpredictable lateral communication, the adhocracy is extremely coordination intensive. New media, such as electronic mail, computer conferencing and electronic bulletin boards, can make the coordination easier and, therefore, enable the adhocracy to work much more effectively. Computer networks can help find and coordinate people with diverse knowledge and skills from many parts of an organization.

Moreover, computer-based technologies can transfer information not only faster and more cheaply but also more selectively. These capabilities help to mitigate information overload. Systems now exist to help people find, filter and sort their electronic mail based on topic, sender and other characteristics. Together these new coordination technologies can speed up the "information metabolism" of organizations—the rate at which firms can take in, move, digest and respond to data.

Abundant information poses two potential difficulties for organizational power. Some people worry that managers may become "Big Brothers" who use the information to exert stronger centralized control over those who work for them. Others fear that if power is decentralized throughout the organization, workers might use their newfound power to serve their own narrow interests, leading to organizational chaos.

In fact, neither dark vision has been realized. Instead what appears to be happening is a paradoxical combination of centralization and decentraliztion. Because information can be distributed more easily, people lower in the organization can now become well enough informed to make more decisions more effectively. At the same time, upper-level managers can more easily review decisions made at lower levels. Thus, because lower-level decision makers know they are subject to spot-checking, senior managers can retain or even increase their central control over decisions.

The changes at Phillips Petroleum Company illustrate this process. Previously, senior managers decided what price to set for petroleum products. These critical decisions depended on the recommendations of staff analysts several levels down. When Phillips Petroleum developed an executive information system, senior managers began to make some of these decisions more directly based on the global information provided by the system. The senior executives soon realized, however, that they could pass on this global information directly to local terminal managers, who could take into account information such as competitors' prices. By decentralizing the pricing decision in this way, the company made sounder, more profitable pricing strategies in each area of the country.

Another way of understanding this

paradoxical effect is to realize that new technology does not just redistribute power. It can provide a sense of more power for everyone. For example, the agents of several insurance companies currently carry laptop computers when they visit the homes of customers. The agents use the computers to fill out applications and to project premiums and benefits. But typically, underwriters at the corporate headquarters require several weeks to review the applications and to issue new policies.

Soon the underwriting rules for certain routine policies will be included in the laptop computer itself. The agent will be able to issue these policies immediately in the customers' homes.

These systems will thus "empower" the agents, who will control the time and place of the policy-acceptance decisions and can make sales immediately. The authority of the central underwriters will increase as well, because the rules they have created will be applied consistently. The underwriters will also be able to devote more time to analyzing interesting and potentially more profitable nonroutine cases.

Information technology not only changes power; it also changes time. On the one hand, time has expanded. Electronic mail, voice mail and facsimile transmissions can be sent or received at any time of day or night, almost anywhere around the globe. Similarly, customers of automatic teller machines and some stock markets can make transactions around the clock. The "work day" has much less meaning, and companies can compete by expanding the times their services are available.

On the other hand, time has contracted. Companies can also compete on speed. For instance, effective coordination can reduce the time needed to develop new products, deliver orders or react to customer requests. Management teams, such as the one at Phillips Petroleum, have information available throughout the management hierarchy, which enables them to react to market conditions much more quickly. Decisions that might have taken days in the past can now be made within hours or minutes.

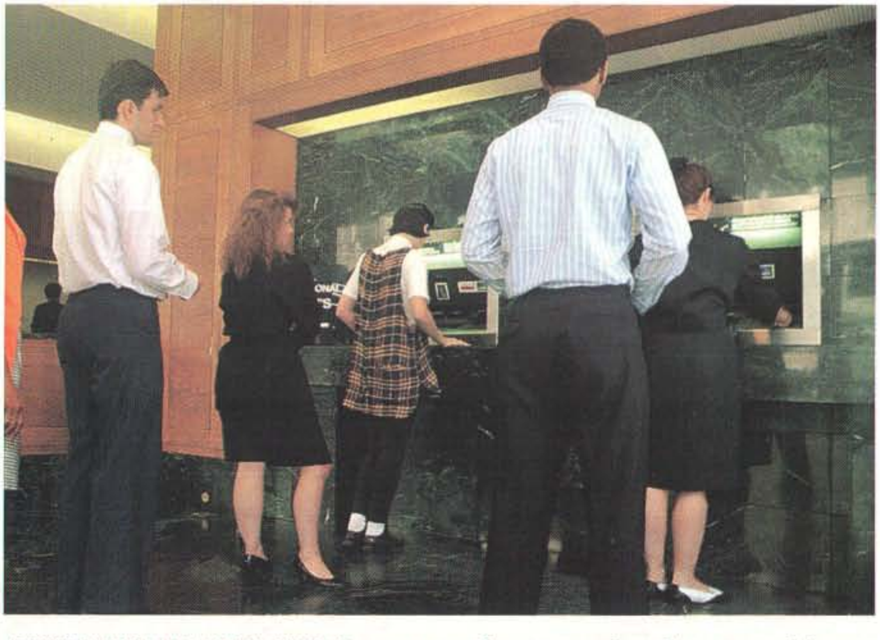
The changes discussed so far require no great predictive leaps; they are already happening. What will happen as information technology improves even more? What other kinds of organizations might emerge in the globally interconnected world that the technologies make feasible?

One possibility is the increasing importance of "answer networks," networks of experts available to answer questions in different areas. One might go to these services with questions such as "How many bars of soap were sold in Guatemala last year?" or "What are the prospects of room-temperature superconductivity in consumer products by 1995?" The services would include massive data bases and layers of human experts in many different topic areas. Some questions will be easily answerable from information in a data base. Others will be referred to progressively more knowledgeable human experts. Depending on how much one is willing to spend and how quickly one wants the answer, the response might range from a newspaper clipping to a personal reply from a Nobel laureate scientist. Similar but limited services exist today—product hot lines and library reference desks are examples—but computer networks and data bases will make such services much less expensive, much more valuable and, therefore, much more widely used.

Electronically mediated markets can also assemble armies of "intellectual mercenaries" virtually overnight. For instance, there may be a large number of consultants who make their living doing short-term projects over the network. If a manager has a job to be done, such as evaluating a loan or designing a lawnmower, he or she could quickly assemble a team by advertising electronically or by consulting a data base of available people. The data base might contain not only the skills and billing rates of prospective workers but also unedited comments from others who had used their services before. Although consulting firms and advertising agencies sometimes work like this now, pervasive networks will allow teams to be assembled much more quickly, for shorter projects and from many different organizations.

This kind of market for services might be used inside an organization as well. Instead of always relying on supervisors to allocate the time of people who work for them, extensive internal markets for the services of people and groups may exist. Murray Turoff of the New Jersey Institute of Technology has suggested how such a system might work. Someone with a short programming project to be done, for instance, might advertise internally for a programmer. Bids and payments for this internal market could be in real dollars or some kind of point system. The bids from programmers would indicate their skill and availability. The payments that programmers receive would reflect how valuable they had been to other parts of the organization.

Improved technology can also help create decision-making structures that integrate qualitative input from many people. For instance, in making complex decisions, such as where to locate a new plant, the amassing of many facts and opinions is critical. Today companies often make such decisions after incomplete discussions with only a few of the people whose knowledge



AUTOMATIC TELLER MACHINES, once a novelty, are now largely a competitive necessity; single innovations in information technology seldom provide long-term advantages. To remain competitive, firms must keep innovating to stay a step ahead or use existing technology to enhance some other advantage.

SUBURBS represent a third-order effect of improved transportation technology: with cheap and convenient transportation came "transportation-intensive" structures. Their existence, however, also depended on the American values placed on the importance of home ownership and the moral superiority of rural life. Similarly, the kinds of coordination-intensive structures that will emerge when information becomes freely accessible will depend on the values society finds important. Will the ideals that shaped Boca Raton, Fla., also play a role in future corporate structures?

or point of view might be valuable. In the future, companies may use computer networks to organize and record the issues, alternatives, arguments and counterarguments in graphical form. Then many different people can review and critique the parts of the argument about which they know or care.

For instance, someone in a remote part of the firm might know about plans for a new highway that completely change the desirability of a proposed plant location. As such information accumulates, people can vote on the plausibility of different claims. Then, using all the information displayed in the system, a single person or group can ultimately make the decision.

What will happen as the globally networked society leads to a world in which vast amounts of information are freely available or easily purchased? Clearly, this world will require services, both automated and human, to filter the tremendous amount of information available. In general, as the amount of information increases, people who can creatively analyze, edit and act on information in ways that cannot be automated will become even more valuable.

But what else people will do will depend on the values that are important to them. When trains and automobiles reduced the constraints of travel time, other values became more significant in determining working and living patterns. As Kenneth T. Jackson of Columbia University has documented, for example, American values about the importance of owning one's home and the moral superiority of rural life played a large role in determining the nature of suburbs in the U.S.

Similarly, when the costs of information and coordination are not a barrier to fulfilling people's needs and wants, other values may emerge to shape the workplace and society. The new information technologies will almost certainly help gratify some obvious wants, such as the desire for money. Some of the emerging corporate structures may be especially good at satisfying nonmaterial needs, such as those for challenge and autonomy.

But perhaps these desires are themselves manifestations of some still deeper needs. Psychologists sometimes refer to a need for self-actualization. Others might call this a desire for spiritual fulfillment. To use the new technologies wisely, we will need to think more carefully about what we truly value and how the technology can help us reach our deeper goals.

FURTHER READING

ELECTRONIC MARKETS AND ELECTRONIC HIERARCHIES. Thomas W. Malone, Joanne Yates and Robert I. Benjamin in *Communications of the ACM,* Vol. 30, No. 6, pages 484–497; June 1987.

BEYOND VERTICAL INTEGRATION: THE RISE OF THE VALUE-ADDING PARTNERSHIP. Russell Johnston and Paul R. Lawrence in *Harvard Business Review,* Vol. 66, No. 4, pages 94–104; July–August 1988.

EVALUATION OF STRATEGIC INVESTMENTS IN INFORMATION TECHNOLOGY. Eric K. Clemons in *Communications of the ACM,* Vol. 34, No. 1, pages 22–36; January 1991.

GLOBALIZATION, TECHNOLOGY, AND COMPETITION: THE FUSION OF COMPUTERS AND TELECOMMUNICATIONS IN THE 1990s. Edited by S. P. Bradley, J. A. Hausman and R. L. Noland. Harvard Business School Press, 1993.

Computers, Networks and Education

Globally networked, easy-to-use computers can enhance learning, but only within an educational environment that encourages students to question "facts" and seek challenges

by Alan C. Kay

The physicist Murray Gell-Mann has remarked that education in the 20th century is like being taken to the world's greatest restaurant and being fed the menu. He meant that representations of ideas have replaced the ideas themselves; students are taught superficially about great discoveries instead of being helped to learn deeply for themselves.

In the near future, all the representations that human beings have invented will be instantly accessible anywhere in the world on intimate, notebook-size computers. But will we be able to get from the menu to the food? Or will we no longer understand the difference between the two? Worse, will we lose even the ability to read the menu and be satisfied just to recognize that it is one?

There has always been confusion between carriers and contents. Pianists know that music is not in the piano. It begins inside human beings as special urges to communicate feelings. But many children are forced to "take piano" before their musical impulses develop; then they turn away from music for life. The piano at its best can only be an amplifier of existing feelings, bringing forth multiple notes in harmony and polyphony that the unaided voice cannot produce.

ALAN C. KAY has been a fellow of Apple Computer Inc. since 1984. Before joining Apple, he was a founder and fellow of the Xerox Palo Alto Research Center and, later, chief scientist of Atari. One of the pioneers of personal computing, he is the original designer of the overlapping-window user interface and Smalltalk, the first completely object-oriented language. Kay has worked with children for most of his career because, he says, "the media that powerfully shape our ways of thinking must be made accessible as early in life as possible." His interests outside of computing include musical performance and instrument design and "trying to learn more about the world in which we find ourselves." He also plays tennis whenever he can.

The computer is the greatest "piano" ever invented, for it is the master carrier of representations of every kind. Now there is a rush to have people, especially schoolchildren, "take computer." Computers can amplify yearnings in ways even more profound than can musical instruments. But if teachers do not nourish the romance of learning and expressing, any external mandate for a new "literacy" becomes as much a crushing burden as being forced to perform Beethoven's sonatas while having no sense of their beauty. Instant access to the world's information will probably have an effect opposite to what is hoped: students will become numb instead of enlightened.

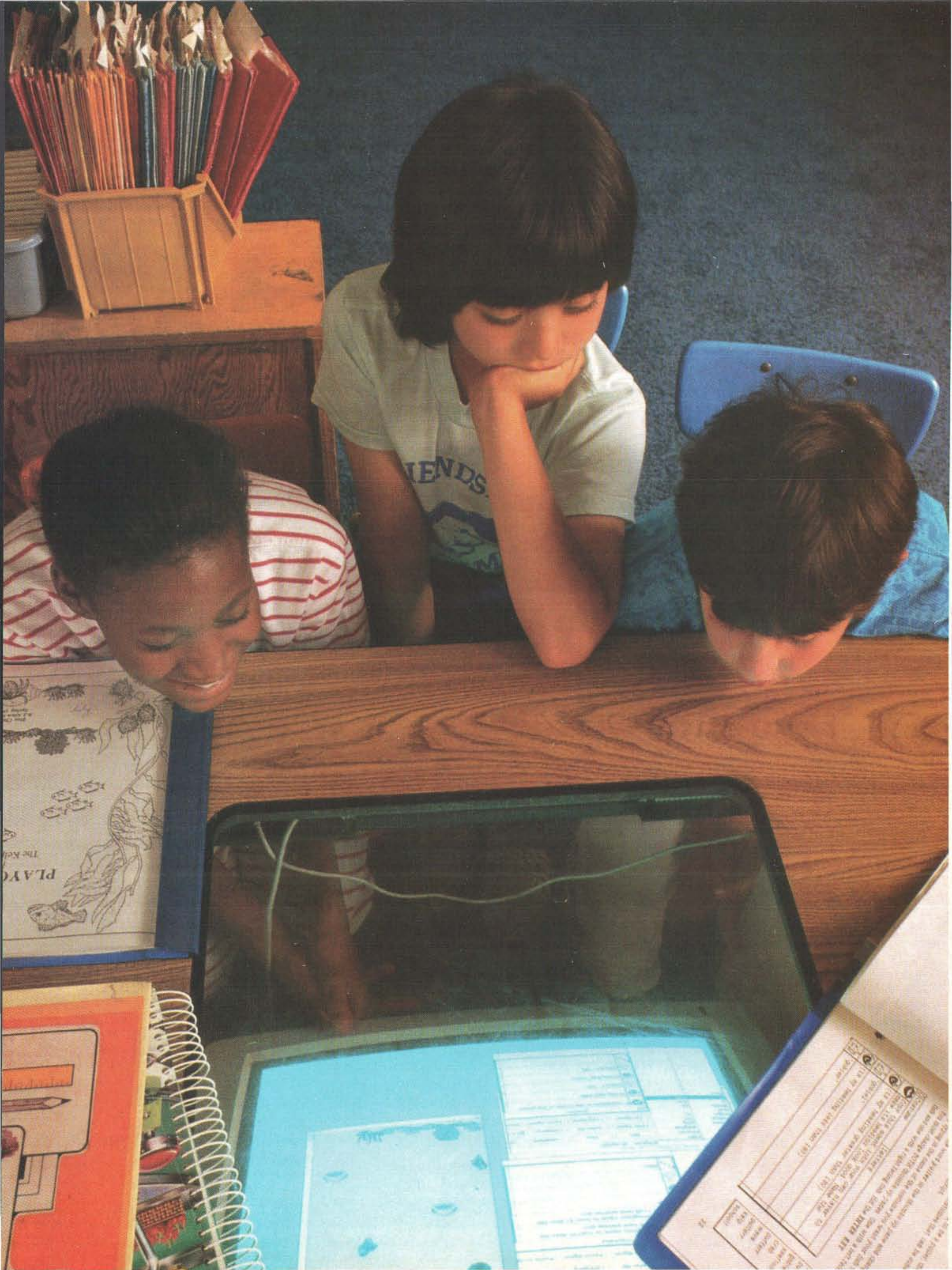
In addition to the notion that the mere presence of computers will improve learning, several other misconceptions about learning often hinder modern education. Stronger ideas need to replace them before any teaching aid, be it a computer or pencil and paper, will be of most service. One misconception might be called the fluidic theory of education: students are empty vessels that must be given knowledge drop by drop from the full teacher-vessel. A related idea is that education is a bitter pill that can be made palatable only by sugarcoating—a view that misses the deep joy brought by learning itself.

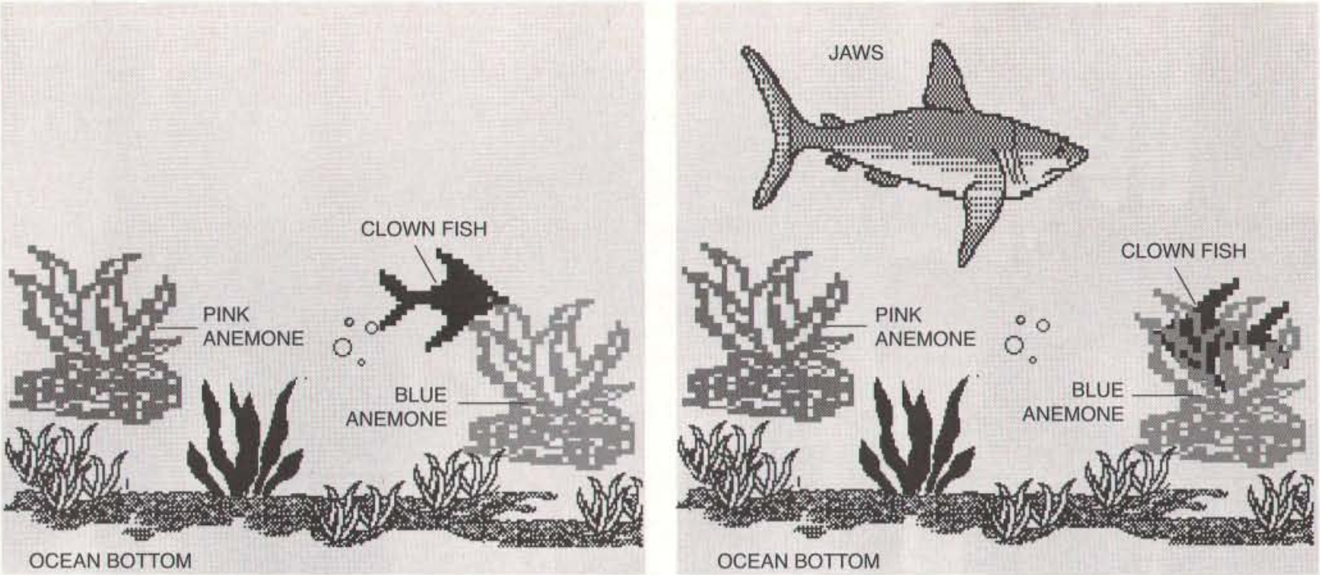
Another mistaken view holds that humans, like other animals, have to make do only with nature's mental bricks, or innate ways of thinking, in the construction of our minds. Equally worrisome is the naive idea that reality is solely what the senses reveal. Finally, and perhaps most misguided, is the view that the mind is unitary, that it has a seamless "I"-ness.

Quite the contrary. Minds are far from unitary: they consist of a patchwork of different mentalities. Jerome S.

STUDENTS at the Open School: Center for Individualization, in Los Angeles, are creating a dynamic simulation of ocean life (*right*) and doing math (*above*) with the help of Macintosh computers, which are set unobtrusively into the desks. In the Open School, which already had a strong curriculum before it obtained computers, the machines do not substitute for teachers. They are thought of as "just another material," like books, paints and clay, that can support the children's activities. In the next few years, notebook-size computers are expected to become available; then children will be able to carry their computers anywhere they go.

Reprinted from the September 1991 issue





CLOWN FISH IS FEATURED in an ocean simulation constructed by nine- and 10-year-olds at the Open School. The fish repeatedly brushes up against an individual sea anemone (*left panel*) to build immunity to its poisonous stings. After immunity is established (*right panel*), the fish can take refuge among the anemone's tentacles whenever a predator (here a shark named Jaws) is near. By constructing simulations, the children learn more about the challenges of being a clown fish and the benefits of symbiosis than they would if they engaged only in more passive activities—such as reading books and observing a fish tank. The author argues that adults, too, learn best when they can test ideas through simulation.

Bruner of New York University has suggested that we have a number of ways to know and think about the world, including doing, seeing and manipulating symbols. What is more, each of us has to construct our own version of reality by main force, literally to make ourselves. And we are quite capable of devising new mental bricks, new ways of thinking, that can enormously expand the understandings we can attain. The bricks we develop become new technologies for thinking.

Many of the most valuable structures devised from our newer bricks may require considerable effort to acquire. Music, mathematics, science and human rights are just a few of the systems of thought that must be built up layer by layer and integrated. Although understanding or creating such constructions is difficult, the need for struggle should not be grounds for avoidance. Difficulty should be sought out, as a spur to delving more deeply into an interesting area. An educational system that tries to make everything easy and pleasurable will prevent much important learning from happening.

It is also important to realize that many systems of thought, particularly those in science, are quite at odds with common sense. As the writer Susan Sontag once said, "All understanding begins with our not accepting the world as it appears." Most science, in fact, is quite literally non-sense. This idea became strikingly obvious when such instruments as the telescope and microscope revealed that the universe consists of much that is outside the reach of our naive reality.

Humans are predisposed by biology to live in the barbarism of the deep past. Only by an effort of will and through use of our invented representations can we bring ourselves into the present and peek into the future. Our educational systems must find ways to help children meet that challenge.

In the past few decades the task before children—before all of us—has become harder. Change has accelerated so rapidly that what one generation learns in childhood no longer applies 20 years later in adulthood. In other words, each generation must be able to quickly learn new paradigms, or ways of viewing the world; the old ways do not remain usable for long. Even scientists have problems making such transitions. As Thomas S. Kuhn notes dryly in *The Structure of Scientific Revolutions*, a paradigm shift takes about 25 years to occur—because the original defenders have to die off.

Much of the learning that will go on in the future will necessarily be concerned with complexity. On one hand, humans strive to make the complex more simple; categories in language and universal theories in science have emerged from such efforts. On the other hand, we also need to appreciate that many apparently simple situations are actually complex, and we have to be able to view situations in their larger contexts. For example, burning down parts of a rain forest might be the most obvious way to get arable land, but the environmental effects suggest that burning is not the best solution for humankind.

Up to now, the contexts that give meaning and limitation to our various knowledges have been all but invisible. To make contexts visible, make them objects of discourse and make them explicitly reshapable and inventable are strong aspirations very much in harmony with the pressing needs and onrushing changes of our own time. It is therefore the duty of a well-conceived environment for learning to be contentious and even disturbing, seek contrasts rather than absolutes, aim for quality over quantity and acknowledge the need for will and effort. I do not think it goes too far to say that these requirements are at odds with the prevailing values in American life today.

If the music is not in the "piano," to what use should media be put, in the classroom and elsewhere? Part of the answer depends on knowing the pitfalls of existing media.

It is not what is in front of us that counts in our books, televisions and computers but what gets into our heads and why we want to learn it. Yet as Marshall H. McLuhan, the philosopher of communications, has pointed out, the form is much of what does get into our heads; we become what we behold. The form of the carrier of information is not neutral; it both dictates the kind of information conveyed and affects thinking processes.

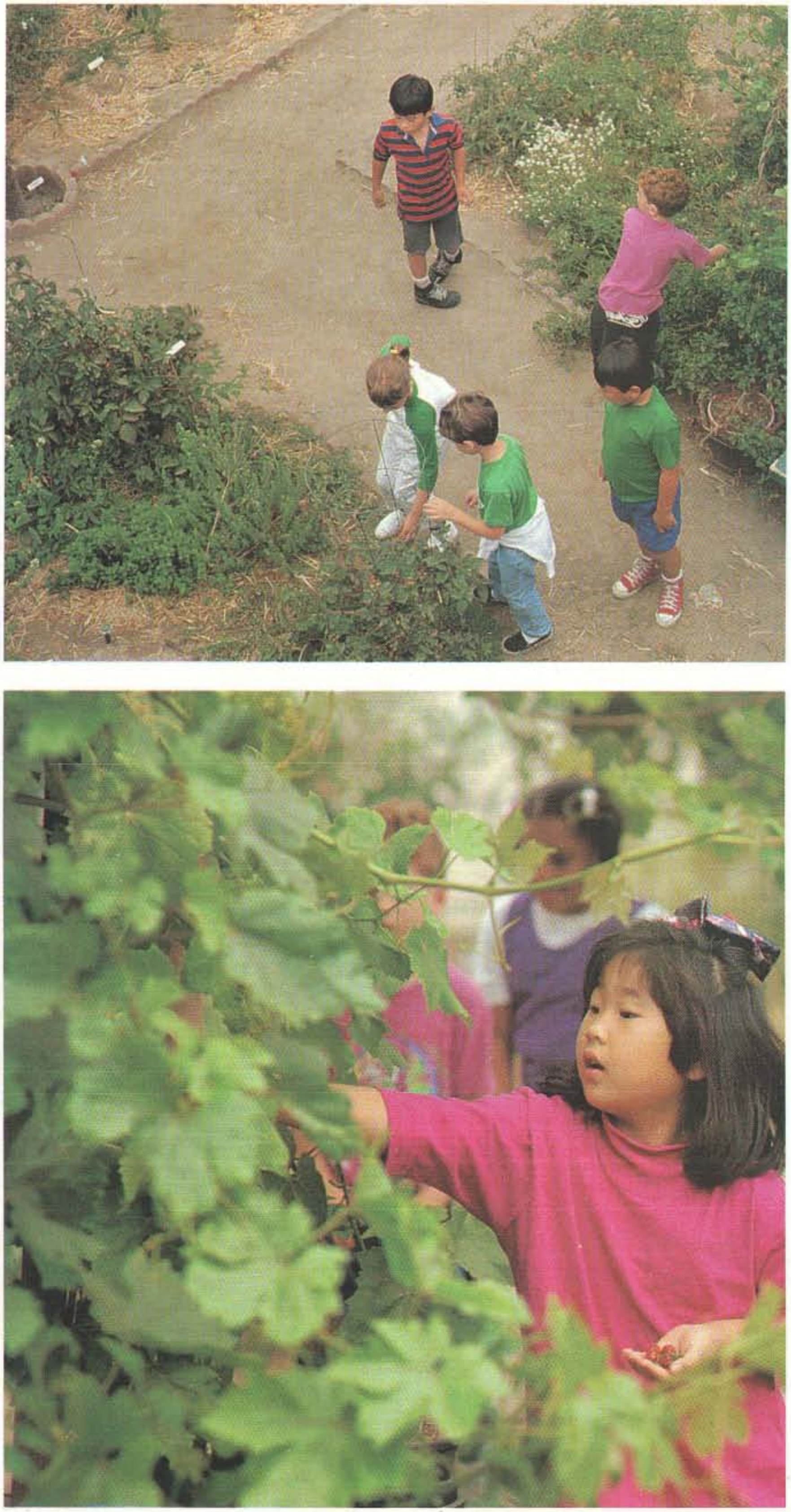
This property applies to all media, not just the new high-tech ones. Socrates complained about writing. He felt it forced one to follow an argument rather than participate in it, and he disliked both its alienation and its persistence. He was unsettled by the idea that a manuscript traveled without the author, with whom no argument was possible. Worse, the author could die and never be talked away from the position taken in the writing.

Users of media need to be aware, too, that technology often forces us to choose between quality and convenience. Compare the emotions evoked by great paintings and illuminated manuscripts with those evoked by excellent photographs of the originals. The feelings are quite different. For the majority of people who cannot make such comparisons directly, there is an understandable tendency to accept the substitution as though nothing were lost. Consequently, little protest has been made over replacing high-resolution photographs of great art (which themselves do not capture the real thing) with lower-resolution videodisc images (which distort both light and space even further). The result is that recognition, not reverie, is the main goal in life and also in school, where recognition is the highest act to which most students are asked to aspire.

When convenience is valued over quality in education, we are led directly to "junk" learning. This is quite analogous to other junk phenomena, pale substitutions masquerading for the real thing. Junk learning leads to junk living. As Neil M. Postman of New York University says, whether a medium carries junk is not important, since all media have junk possibilities. But one needs to be sure that media incapable of carrying important kinds of discourse—for example, television—do not displace those that can.

Media can also lure us into thinking we are creating by design when in fact we are just tinkering. Consider the difficulty of transforming clay—a perfectly malleable and responsive substance—into anything aesthetically satisfying. Perfect "debugability," or malleability, does not make up for lack of an internal image and shaping skills. Unfortunately, computers lend themselves to such "clay pushing"; they tempt users to try to debug constructions into existence by trial and error.

Finally, as McLuhan noted, the instant communication offered by today's media leads to fragmentation. Sequence and exposition are replaced by isolated,



WALKWAY through a garden (*top*) outside the Open School was designed by the third graders, who chose a herringbone pattern to ensure easy access to all plots. The children settled on the pattern after creating and debating many models, often with the help of their computers. The garden is part of the Life Lab project, in which children plan, plant, tend and enjoy the fruits of their own garden (*bottom*) as a way of learning about the interaction of living things with the environment.

context-free factoids, often presented simply because they are recent. Two hundred years ago the Federalist papers—essays by James Madison, Alexander Hamilton and John Jay arguing for ratification of the U.S. Constitution—were published in newspapers in the 13 colonies. Fifty years later the telegraph and its network shifted the goals of news from depth to currency, and the newspapers changed in response. Approximately 100 years after that, television started shifting the emphasis of news from currency to visual immediacy.

Computers have the same drawbacks as other media, and yet they also offer opportunities for counteracting the inherent deficits. Where would the authors of the Constitution publish the Federalist papers today? Not in a book; not enough people read books. Not in newspapers; each essay is too long. Not on the television; it cannot deal with thoughtful content. On computer networks? Well, computer displays, though getting better every year, are not good enough for reading extended prose; the tendency is to show pictures, diagrams and short "bumper sticker" sentences, because that is what displays do well.

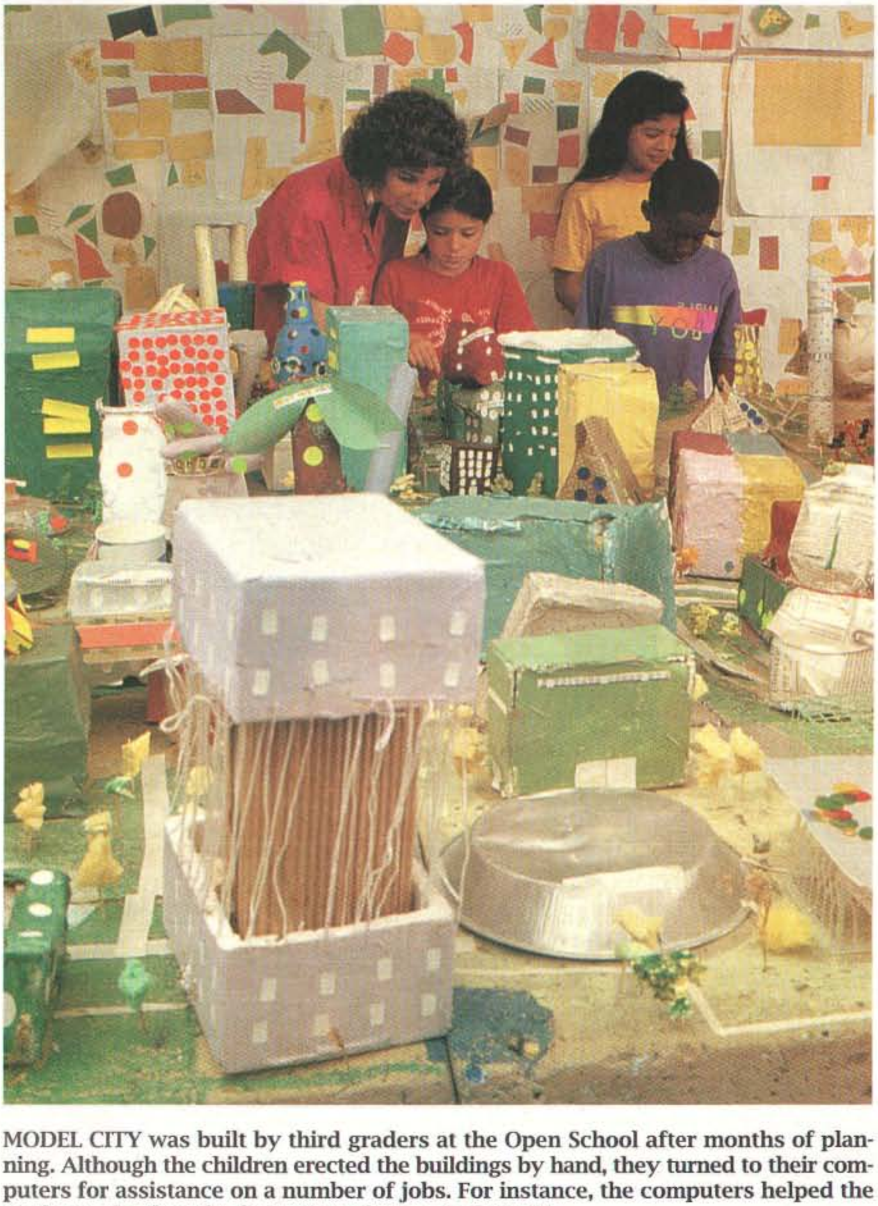
But the late 20th century provides an interesting answer to the question: transmitting over computer networks a simulation of the proposed structure and processes of the new Constitution. The receivers not only could run the model but also could change assumptions and even the model itself to test the ideas. The model could be hyperlinked to the sources of the design, such as the constitution of Virginia, so that "readers" might readily compare the new ideas against the old. (Hyperlinking extends any document to include related information from many diverse sources.) Now the receivers would have something stronger than static essays. And feedback about the proposals—again by network—could be timely and relevant.

Five years ago, intent on studying firsthand the strengths and weaknesses of computers as amplifiers for learning, my colleague Ann Marion and I, in collaboration with the Open School: Center for Individualization, in Los Angeles, set up a research project called the Apple Vivarium Program. We and the principal, Roberta Blatt, were not trying to improve the already excellent school by introducing technology. We were trying to better understand the value computers might have as supporting media.

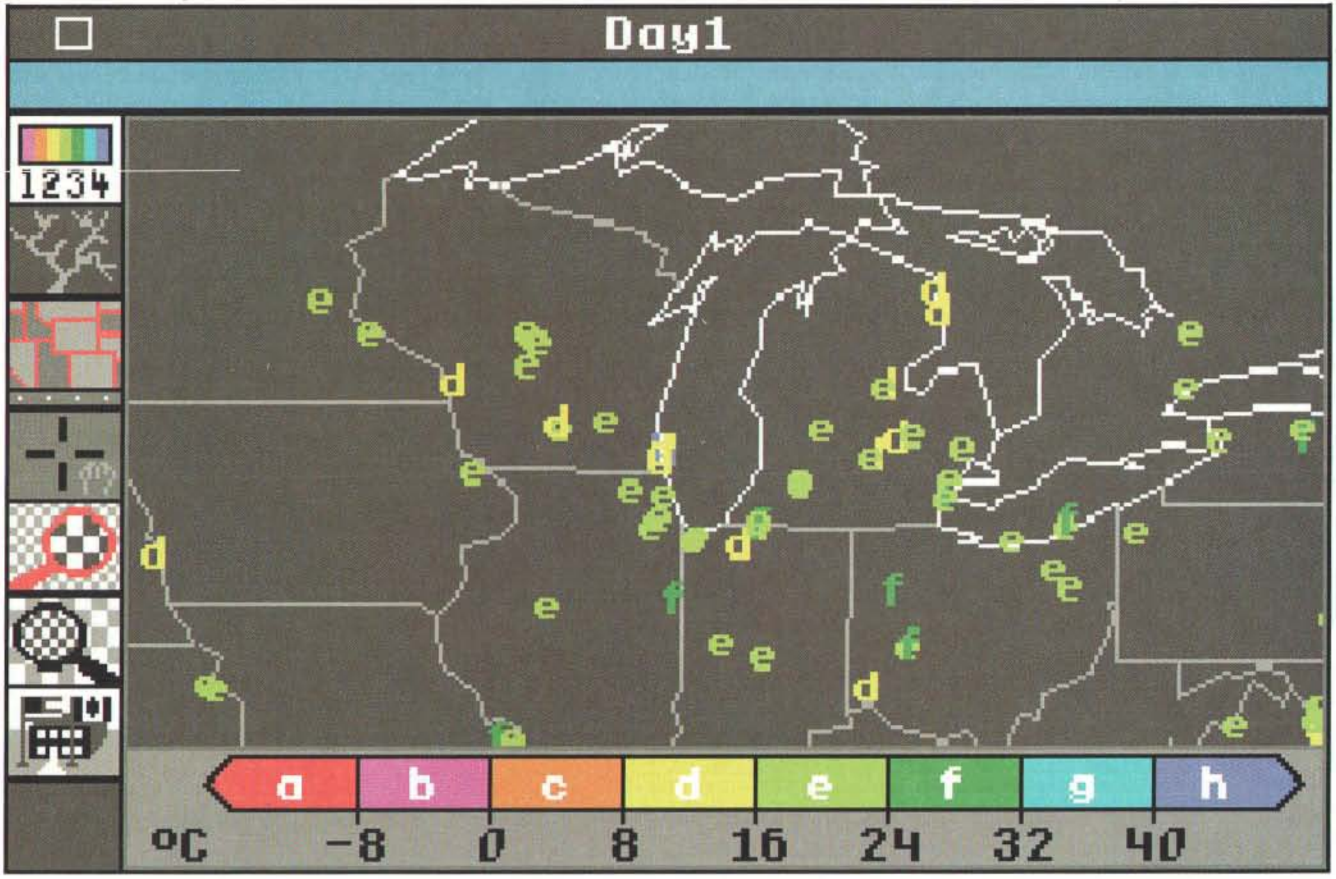
Children are bused in and, as is the case with other busing schools in Los Angeles, are selected by lot so that the racial balance is roughly in accord with that of the city as a whole. Parents have to be interested enough in their children and the school's teaching approach to put their children on the list for consideration. Parental interest and involvement are key factors that have made the school a success. One could even argue that the educational approach in a classroom is not nearly as important as the set of values about learning found in the home. If those exist, almost any process will work, although some may be more enjoyable and enriching than others.

We particularly wanted to investigate how children can be helped to understand that animals, people and situations are parts of larger systems that influence one another. We therefore focused much of our work on the study of biology and ecology. Studies of the design and functioning of large cities also give children an awareness of such complexity. Doreen Nelson of the California Polytechnic Institute has been teaching city design to children for many years; on the basis of her work, our study group introduced a large-scale city-building project for the third graders. We also helped the school develop a major theater program, so the children might see how art and systems work from the inside.

What does it mean to learn about biology as it relates to us and our world? All creatures consist of and are part of many systems that range from the molecular to the planetary. A weak way to approach this romance—in which we are at once part of the scenery, bit



MODEL CITY was built by third graders at the Open School after months of planning. Although the children erected the buildings by hand, they turned to their computers for assistance on a number of jobs. For instance, the computers helped the students simulate the formation of smog in their city.



PLOT OF TEMPERATURES in the Great Lakes region of the U.S. is part of an international map created from data collected by students in hundreds of schools. The children took measurements at the same time of day and pooled them through a network. Making such maps is part of the *Weather in Action* unit of the National Geographic Kids Network curriculum. It is also an example of how networks can enhance scientific collaboration—for children as much as for adults.

players, star-crossed lovers, heroes and villains—would be a well-meaning attempt to use books, computers or other representational media as "delivery vehicles." There could be videodiscs showing plant and animal growth, and the students could have network access to data about crop yields, taxonomies of animals and plants, and so forth. But why substitute a "music appreciation" approach for the excitement of direct play? Why teach "science and math appreciation," when the children can more happily (and to better effect) actually create whole worlds?

What is great in biology and humankind's other grand investigations cannot be "delivered." But it can be learned—by giving students direct contact with "the great chain of being," so that they can internally generate the structures needed to hold powerful ideas. Media of all kinds can now be used to amplify the learning experience, whereas before they acted as a barrier to the "good stuff."

The Open School is nothing if not straightforward. Because "things that grow" is the essence of what is called the Life Lab program, the children made a garden, tearing up part of their asphalt playground to get good clean dirt. The third graders, while in the midst of their city-building project, spent months modeling and debating designs for the garden. They ultimately arrived at a practical, child-scaled pattern featuring a herringbone-shaped walkway that puts every plot in reach.

Not surprisingly, the children found that the simulation capability of their computers helped them examine the merits of many different walkway designs. Like modern-day architects, they used the computer to help construct models of their ideas. Teachers Dolores Patton and Leslie Barclay facilitated the process, but it was the children who came up with the ideas.

There are many Life Lab schools in California. Because they are engaged in similar pursuits, they have things to say to one another. For them, networks serve as much more than a conduit for retrieving fixed data; they allow students to develop knowledge of their own collaboratively. For example, it is easy to make one's own weather maps on the basis of simultaneous recordings of temperature and barometric pressure and the like and to argue via network about what the maps mean.

Computer animation can be used to ponder the patterns more readily. A fairly easy inference is that pressure changes seem to go from west to east. Could this have anything to do with the rotation of the earth? The directions of winds are more complicated, since they are more affected by features of the terrain. Do they match up with pressure changes?

We can go still deeper. Children are capable of much depth and attention to quality when they are thinking about questions that seem important to them. Why do animals do what they do? Why do humans do what we do? These are vital issues. Close observation, theories and role-playing help. Reading books about animal behavior helps. The teacher can even explain some ideas of the Nobel laureate Niko Tinbergen, such as the suggestion that animal behavior is organized into modules of innate patterns.

But these are just words. Now the children can make dynamic models of animal behavior patterns to test Tinbergen's concepts themselves.

Can nine- and 10-year-old children actually capture and understand the mentality of a complex organism, such as a fish? Teacher B. J. Allen-Conn spent several summers learning about intricate ecological relations in the oceans. She searched for ways to express how an individual's behavior is altered by interactions with many other animals. At the same time, Michael Travers, a graduate student from the Media Laboratory at the Massachusetts Institute of Technology who was working with us, built several animal simulations, among them fish behaviors described by Tinbergen. Then Scott Wallace and others in our group turned these various ideas into Playground, a simulation construction kit for children.

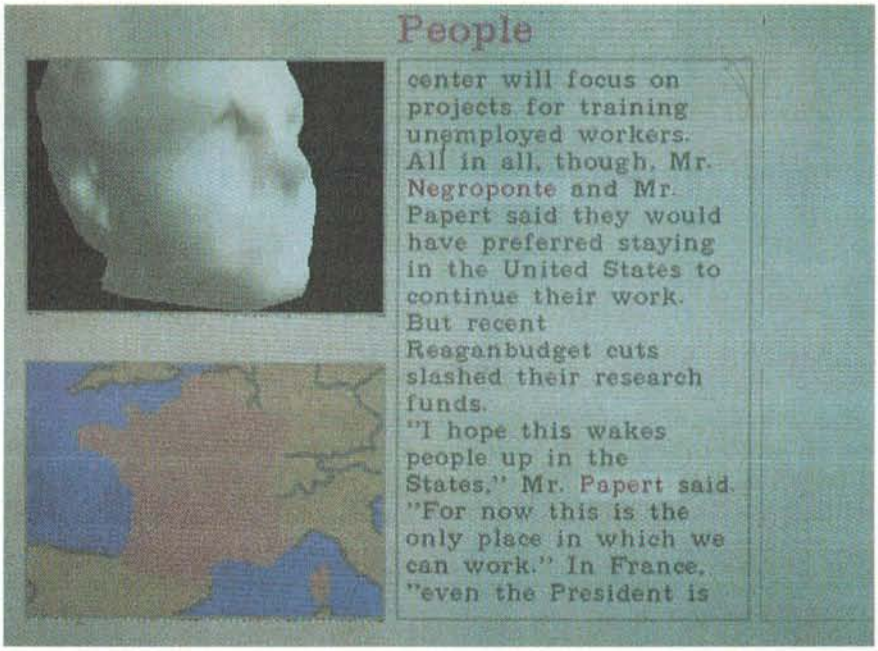
Children are particularly enthralled by the clown fish, which exhibits all expected fishlike behaviors (such as feeding, mating and fleeing from predators) but also displays a fascinating way of protecting itself. It chooses a single sea anemone and gradually acclimates to the anemone's poison over a period of several days. When acclimation is complete, the clown fish has a safe haven where it can hide if a predator comes hunting.

It is fairly easy to build a simple behavior in Playground, and so the children produce simulations that reflect how the fish acts when it gets hungry, seeks food, acclimates to an anemone and escapes from predators. Later they can explore what happens when scripts conflict. What happens if the animal is very hungry yet there is a predator near the food? If the animal is hungry enough, will it start eyeing the predator as possible food? Do the fish as a group fare best when each animal is out for itself, or does a touch of altruism help the species overall?

For an adult, the children's work would be called Artificial Intelligence Programming Using a Rule-Based Expert Systems Language. We researchers and the teachers and children see the dynamic simulations as a way of finding out whether theories of animal behavior apply to the real world.

Computers in the Open School are not rescuing the school from a weak curriculum, any more than putting pianos in every classroom would rescue a flawed music program. Wonderful learning can occur without computers or even paper. But once the teachers and children are enfranchised as explorers, computers, like pianos, can serve as powerful amplifiers, extending the reach and depth of the learners.

Many educators have been slow to recognize this concept of knowledge ownership and to realize that children, like adults, have a psychological need for a personal franchise in the culture's knowledge base. Most schools force students to learn somebody else's knowledge. Yet, as John Holt, the teacher and philosopher of education, once said, mathematics and science would probably be learned better if they were made illegal. Children learn in the same way as adults, in that they learn best when they can ask their own questions, seek answers in many places, consider different perspectives, exchange views with others and add their own findings to existing understandings.



People

center will focus on
projects for training
unemployed workers.
All in all, though, Mr.
Negroponte and Mr.
Papert said they would
have preferred staying
in the United States to
continue their work.
But recent
Reaganbudget cuts
slashed their research
funds.
"I hope this wakes
people up in the
States," Mr. Papert said.
"For now this is the
only place in which we
can work." In France,
"even the President is

EXCERPT FROM A NEWSCLIP is part of a newspaper tailored to the interests of a single individual; the text was produced several years ago by the software program NewsPeek, which the author and Walter Bender designed when they were at the Massachusetts Institute of Technology. The program is an early prototype for one kind of "agent," a system that can learn a user's goals and retrieve relevant information on the person's behalf. Such agents will one day be essential for navigating through the mass of information that will be available on networks.

Ten years from now, powerful, intimate computers will become as ubiquitous as television and will be connected to interlinked networks that span the globe more comprehensively than telephones do today. My group's experience with the Open School has given us insight into the potential benefits of this technology for facilitating learning.

The first benefit is great interactivity. Initially the computers will be reactive, like a musical instrument, as they are today. Soon they will take initiatives as well, behaving like a personal assistant. Computers can be fitted to every sense. For instance, there can be displays for vision; pointing devices and keyboards for responding to gesture; speakers, piano-type keyboards and microphones for sound—even television cameras to recognize and respond to the user's facial expressions. Some displays will be worn as magic glasses and force-feedback gloves that together create a virtual reality, putting the user inside the computer to see and touch this new world. The surface of an enzyme can be felt as it catalyzes a reaction between two amino acids; relativistic distortions can be directly experienced by turning the user into an electron traveling at close to the speed of light.

A second value is the ability of the computers to become any and all existing media, including books and musical instruments. This feature means people will be able (and now be required) to choose the kinds of media through which they want to receive and communicate ideas. Constructions such as texts, images, sounds and movies, which have been almost intractable in conventional media, are now manipulatable by word processors, desktop publishing, and illustrative and multimedia systems.

Third, and more important, information can be presented from many different perspectives. Marvin L. Minsky of M.I.T. likes to say that you do not understand anything until you understand it in more than one way. Computers can be programmed so that "facts" retrieved in one window on a screen will automatically cause supporting and opposing arguments to be retrieved in a halo of surrounding windows. An idea can be shown in prose, as an image, viewed from the back and the front, inside or out. Important concepts from many different sources can be collected in one place.

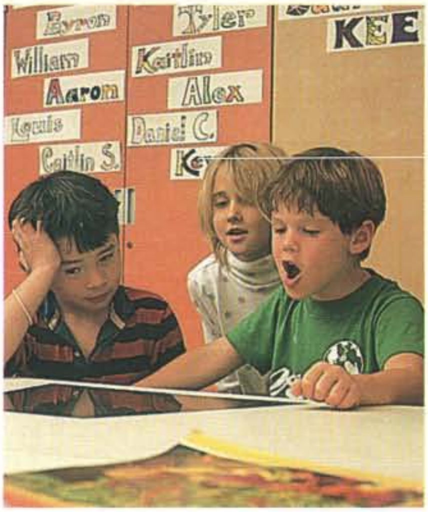
Fourth, the heart of computing is building a dynamic model of an idea through simulation. Computers can go beyond static representations that can at best argue; they can deliver sprightly simulations that portray and test conflicting theories. The ability to "see" with these stronger representations of the world will be as important an advance as was the transition to language, mathematics and science from images and common sense.

A fifth benefit is that computers can be engineered to be reflective. The model-building capabilities of the computer should enable mindlike processes to be built and should allow designers to create flexible "agents." These agents will take on their owner's goals, confer about strategies (asking questions of users as well as answering their queries) and, by reasoning, fabricate goals of their own.

Finally, pervasively networked computers will soon become a universal library, the age-old dream of those who love knowledge. Resources now beyond individual means, such as supercomputers for heavy-duty simulation, satellites and huge compilations of data, will be potentially accessible to anyone.

For children, the enfranchising effects of these benefits could be especially exciting. The educator John Dewey noted that urban children in the 20th century can participate only in the form, not the content, of most adult activities; compare the understanding gained by a city girl playing nurse with her doll to that gained by a girl caring for a live calf on a farm. Computers are already helping children to participate in content to some extent. How students from preschool to graduate school use their computers is similar to how computer professionals use theirs. They interact, simulate, contrast and criticize, and they create knowledge to share with others.

When massively interconnected, intimate computers become commonplace, the relation of humans to their information carriers will once again change qualitatively. As ever more information becomes available, much of it conflicting, the ability to critically assess the value and validity of many different points of view and to recognize the contexts out of which they arise will become increasingly crucial. This facility has been extremely important since books became widely available, but making comparisons has been quite difficult. Now comparing should become easier, if people take advantage of the positive values computers offer.



CHILDREN AT A COMPUTER in the Open School are clearly engrossed in their work. If used properly, the author notes, computers can be "powerful amplifiers, extending the reach and depth of the learners."

Computer designers can help as well. Networked computer media will initially substitute convenience for verisimilitude, and quantity and speed for exposition and thoughtfulness. Yet well-designed systems can also retain and expand on the profound ideas of the past, making available revolutionary ways to think about the world. As Postman has pointed out, what is required is a kind of guerilla warfare, not to stamp out new media (or old) but to create a parallel consciousness about media—one that gently whispers the debits and credits of any representation and points the way to the "food."

For example, naive acceptance of onscreen information can be combated by designs that automatically gather both the requested information and instances in which a displayed "fact" does not seem to hold.

An on-line library that retrieves only what it is requested produces tunnel vision and misses the point of libraries; by wandering in the stacks, people inevitably find gems they did not know enough to seek. Software could easily provide for browsing and other serendipitous ventures.

Today facts are often divorced from their original context. This fragmentation can be countered by programs that put separately retrieved ideas into sequences that lead from one thought to the next. And the temptation to "clay push," to create things or collect information by trial and error, can be fought by organizational tools that help people form goals for their searches. If computer users begin with a strong image of what they want to accomplish, they can drive in a fairly straightforward way through their initial construction and rely on subsequent passes to criticize, debug and change.

If the personally owned book was one of the main shapers of the Renaissance notion of the individual, then the pervasively networked computer of the future should shape humans who are healthy skeptics from an early age. Any argument can be tested against the arguments of others and by appeal to simulation. Philip Morrison, a learned physicist, has a fine vision of a skeptical world: "...genuine trust implies the *opportunity* of checking wherever it may be wanted.... That is why it is the evidence, the experience itself and the argument that gives it order, that we need to share with one another, and not just the unsupported final claim."

I have no doubt that as pervasively networked intimate computers become common, many of us will enlarge our points of view. When enough people change, modern culture will once again be transformed, as it was during the Renaissance. But given the current state of educational values, I fear that, just as in the 1500s, great numbers of people will not avail themselves of the opportunity for growth and will be left behind. Can society afford to let that happen again?

FURTHER READING

UNDERSTANDING MEDIA: THE EXTENSIONS OF MAN. Marshall McLuhan. McGraw-Hill, 1965.

TOWARD A THEORY OF INSTRUCTION. Jerome S. Bruner. Harvard University Press, 1966.

THE STRUCTURE OF SCIENTIFIC REVOLUTIONS. Thomas S. Kuhn. University of Chicago Press, 1970.

AMUSING OURSELVES TO DEATH: PUBLIC DISCOURSE IN THE AGE OF SHOW BUSINESS. Neil Postman. Viking Penguin, 1985.

THE RING OF TRUTH: AN INQUIRY INTO HOW WE KNOW WHAT WE KNOW. Philip Morrison and Phylis Morrison. Random House, 1989.

Infrastructure for the Global Village

A high-capacity network will not be built without government investment

by Al Gore

New technologies that enhance the ability to create and understand information have always led to dramatic changes in civilization. The printing press unleashed the forces that led to the birth of the modern nation-state. It made possible the widespread distribution of civic knowledge that enabled the average citizen to affect political decisions.

Now come distributed networks connecting a myriad of computers, ranging from megaflop machines and workstations to desktop and personal laptop units. All are becoming less expensive in each successive generation that emerges. There is no longer any doubt that such machines will reshape human civilization even more quickly and more thoroughly than did the printing press. Gutenberg's invention, which so empowered Jefferson and his colleagues in their fight for democracy, seems to pale before the rise of electronic communications and innovations, from the telegraph to television, to the microprocessor and the emergence of a new computerized world—an information age.

AL GORE was elected vice president of the U.S. in 1992 after representing Tennessee in the U.S. Senate for eight years and a Tennessee congressional district in the House of Representatives for eight years before that. A 1969 graduate of Harvard University, he also attended Vanderbilt University Divinity School and Vanderbilt Law School. In the Senate, Gore was chairman of the subcommittee on Science Technology and Space of the Commerce, Science and Transportation Committee. He was the author of the High Performance Computing Act of 1990, which would create high-speed fiber-optic computer networks, and wrote legislation to encourage the development of high-definition television.

As in the past, original thinking is required to develop a system that will enable our civilization to make the most sensible use of this potent technology. There are valid concerns that computer networks threaten the privacy and personal freedom of individuals. There are equally valid worries that existing laws do not adequately protect the rights and liberties of computer users to express themselves in the new medium.

Even so, such concerns should not—and indeed, probably cannot—halt the sweeping changes that are already taking place. Just as information spewing forth from the printing press was soon protected in democratic states by a fair and workable legal and ethical code, so, too, a new body of common law is being built to cope with the new medium of the computer network. Rather than holding back, the U.S. should lead by building the information infrastructure, essential if all Americans are to gain access to this transforming technology.

The unique way in which the U.S. deals with information has been the real key to our success. Capitalism and representative democracy rely on the freedom of the individual, so these systems operate in a manner similar to the principle behind massively parallel computers. These computers process data not in one central unit but rather in tiny, less powerful units distributed throughout the computer.

Capitalism works on the same principle. People who are free to buy and sell products or services according to their individual calculations of the costs and benefits of each choice process a relatively limited amount of information but do it quickly. When millions of individuals process information simultaneously, the aggregate result is incredibly accurate and efficient decisions.

Similarly, representative democracy relies on the still revolutionary assumption that the best way for a nation to make political decisions is for all its citizens to process the information relevant to their lives and to express their conclusions in free speech and in votes that are combined with those of millions of others to guide the system as a whole. Communism, by contrast, attempted to bring all the information to a large and powerful central processor, which collapsed when it was overwhelmed by ever more complex information.

Computers are providing people all over the U.S. with the ability to process information and to use their creativity to challenge the world around them in ways never before imagined. More and more, computers are replacing telephones; more and more, computers are able to communicate as well as calculate. To realize the full benefit of the information age, we need to take the next step: high-speed networks must be built that tie together millions of computers, providing capabilities that we cannot even imagine. Consider supercomputers. Although they are far more valuable than their weaker predecessors, they cannot be effectively linked together over existing networks.

In laboratories across our country, scientists and researchers armed with powerful supercomputers are using them to design more effective medicines, safer aircraft and automobiles and more efficient products. Scientists studying the earth's environment can produce complex climate models to help understand global warming and other critical atmospheric trends.

Suppose that we had a national "colaboratory" linking teams at different geographic locations so that researchers could work together on projects involving computational science. Suppose

Reprinted from the September 1991 issue

that a scientist or an engineer designing a product could immediately transfer a design to the manufacturer, who in turn could transfer it instantly to computer-operated machines ready to turn the idea into reality. The benefits, both intellectual and economic, would be phenomenal.

There have been advances in networking technology. We have the technical know-how to make networks that would enable a child to come home from school and, instead of playing Nintendo, use something that looks like a video game machine to plug into the Library of Congress. It could be possible to gain access to the most powerful supercomputers in our nation from every PC in America—and to gain access to digital libraries, another key part of the national information infrastructure we need.

It used to be that a nation's transportation infrastructure determined success in international economic competition. Nations with deep-water ports did better than nations unable to exploit the technology of ocean transportation. Canals, railroads, highways, water and sewer systems—the list is a long one—were infrastructure investments found, through experience, to provide tremendous leverage for improving a nation's competitiveness.

Today transportation is less important compared with other factors, such as the ability to move information and to increase the value of this information. At the same time, however, we are confronted with massive and rapidly multiplying quantities of information. We are forced to deal with what I call "exformation"—accumulated surpluses of data that we know exist but are outside our conscious awareness.

Fields of specialization are becoming narrower and narrower as the amount of information continues to explode, doubling every six months in some disciplines. In a sense, we have automated the process of gathering information without enhancing our ability to absorb its meaning.

Our current national information policy resembles the worst aspects of our old agricultural policy, which left grain rotting in thousands of storage silos while people were starving. We have warehouses of unused information "rotting," while critical questions are left unanswered and critical problems are left unsolved. For example, the *Landsat* satellite is capable of taking a complete photograph of the entire earth's surface every two weeks. It has been operating for nearly 20 years. Yet more than 95 percent of those images, which might be invaluable to farmers, environmental scientists, geologists, educators, city planners and businesses, have never been seen by human eyes.

Luckily, just as we confront this fantastic problem of so-called exformation, all these silos filled with "rotting data," we see the appearance of the supercomputer. These machines promise the ability to configure or organize information into a form easy for us to absorb. They can search through vast fields of data to find those bits that have special relevance at the moment.

But first we must build a way to use it effectively. Typically, software development follows hardware development,

NETWORKING SESSION is carried on by Vice President Al Gore at an electronic town meeting early in 1994. The session, organized by CompuServe Information Service, enabled participants to ask the vice president about the information highway.

and policy lags behind both. Yet it is policy that can determine whether we reap the benefits of this new technology. In too many cases, we have mastered the technology but failed to muster the political commitment and the appropriate policies.

Federal policies can make a difference in several key areas of the computer revolution. The U.S. Advanced Research Projects Agency (ARPA), the National Aeronautics and Space Administration, the National Science Foundation and the Department of Energy have all spent millions of dollars developing the next generation of supercomputers. It is safe to say that without federal seed money, Cray Research and other American firms would not now dominate the world market for supercomputers.

The funds from those agencies have also built today's research computer networks, such as NSFNET, and underwritten the development of software for top-of-the-line machines. Additional federal research funding is also needed to develop easy-to-use applications for supercomputers and networks (especially in industry and medicine) and to train people in computational science. The massive amounts of data in our information silos, much of it in electronic form, could be made available to researchers over networks. But resources are needed to put the data in a useful format and to catalogue and distribute them.

Most important, we need a commitment to build the high-speed data highways. Their absence constitutes the largest single barrier to realizing the potential of the information age. Present policy, which is based on copper-wire networks, is hindering deployment of the new fiber technology. There is a policy gridlock—some say a "graphic jam"—because the interests that built and still run our existing infrastructure resist changes that might intensify competition.

To realize the full benefit of the information age, high-speed networks that tie together millions of computers must be built.

Other countries, including not only advanced nations such as Japan and Germany but also many countries in the developing world that are just now building a universal telephone system, do not have this problem. If we do not break the communications gridlock, our foreign competitors could once again reap the benefits of U.S. technology while we remain mired in the past.

The most effective way to break the stalemate would be to show the American people what fiber-optic networks could offer them. Most Americans are only vaguely familiar with even existing networks, yet they rely on networks every day. Everyone, from shoppers at the checkout counter to consumers making a withdrawal from an automated teller machine, is dealing with a computer through a network.

Thousands of researchers now use megabit-per-second networks to communicate with computers a continent away. A few hundred people, exemplified by Robert E. Kahn, president of the Corporation for National Research Initiatives, are building even faster, gigabit networks to link supercomputers, allowing two, three or more supercomputers to work together as one.

GLOBAL CONFERENCE over interactive television took place when Vice President Gore announced GLOBE: Global Learning and Observation to Benefit the Environment. The program involves teachers and students.

Slower networks that operate at 45 megabits a second or less, like NSFNET, are being used by more than half a million researchers, educators and students at colleges, universities and federal and private laboratories to communicate with their colleagues, to access data bases and to collect data from remote instruments. Several high-tech enterprises, such as IBM and AT&T Bell Laboratories, also depend on such networks. Banks and other financial institutions, many factories and other businesses now rely on networks that operate in the range of 50,000 to 100,000 bits per second to exchange electronic mail, financial transactions, stock quotations and a number of other types of information.

Services like Prodigy, GEnie and CompuServe are now in use in almost half a million homes. Children can use "electronic encyclopedias" to finish their term papers, and their parents can keep up with the news, check bank balances and get consumer tips. Hundreds of schools are linked by low-speed (1,200- or 2,400-bit-per-second) networks, such as Kids' Net, that allow students to exchange electronic mail.

Networks thousands or even millions of times faster would offer greater benefits. Today a child can go to a library and use a computer to get the title of a book. A faster network would bring the book to the child at home, pictures and all. But how do we get there? It is estimated that more than $100 billion of private investment would be needed to connect optical fiber to every home, office, factory, school, library and hospital. Unfortunately, the onslaught of technology is causing fierce competition between once separate entities in telecommunications. The telephone companies want to transmit entertainment; the cable carriers are asking to get into the communications business. And businesses are turning to satellite communications carriers or installing private fiber-optic networks.

All are regulated by vague, outdated, conflicting, constantly changing government telecommunications policies. The result is that the private sector hesitates to jump in and make the investment. We face the classic "chicken and egg" dilemma. Because there is no network, there is no apparent demand for the network; because there is no demand, there is no network.

Consider the interstate highway system. It made sense for a postwar America with lots of new automobiles clogging crooked, two-lane roads. But who was going to pay for it? There was no private business that could afford it and no entrepreneur convinced it was worth it. Similarly, the private sector cannot afford to build the high-speed computer network we need and may not even be convinced of its value.

But, like the interstate highway system, once this network is completed, the demand for its use will skyrocket. We are already seeing private and public concerns building feeder networks that, just as state roads feed into the interstate highways, could make the network more effective. More than 2,000 separate networks make up the Internet, the largest national research network. The amount of traffic is growing at the rate of 10 percent a month.

If we do not break the communications gridlock, our competitors could once again reap the benefits while we remain mired in the past.

For more than 15 years, I have been working to change federal policy so that as a nation we will invest in the critical infrastructure of information superhighways. I have fought for funding for the Internet and its successor, the National Research and Education Network (NREN), which will be almost 100 times faster. Its development costs, about $390 million in federal funds over the next five years, represent less than 1 percent of federal research and development expenditures, yet this network could greatly enhance the productivity and value of the other 99 percent of the research and education dollars we spend.

The national network represents a demonstration project that should build public support and influence public policy. I see it as an electronic battering ram to knock down obsolete policies and outdated skepticism. Already the national network as an idea is serving as a catalyst for action by schools and universities, businesses and research centers and even state and local governments.

Adequate federal investment and leadership will give us the power to leverage the millions of dollars into something greater than the sum of its parts. Federal leadership is needed to provide direction and coordination. Federal dollars represent an initial, rather than an ongoing, investment, intended to spark the development that will create the demand for this network as a commercial enterprise.

The "demonstration effect" of the national network will rapidly develop enormous demand for the new information services its high capacity will make possible. It will link the hundreds of computer networks, operated by state and local governments, nonprofit and for-profit corporations, universities and others, to a federally funded "backbone," capable of transmitting billions of bits of data per second.

Without federal funding for this national network, we would end up with a balkanized system, consisting of dozens of incompatible parts. The strength of the national network is that it will not be controlled or run by a single entity. Hundreds of different players will be able to connect their own networks to this one.

The potential benefits of a high-speed network are so obvious and so great that a strong, broad coalition has formed in support. The research and education community, the telecommunications industry, the computer industry and other high-tech industries are all enthusiastic.

But although there is much agreement on the need for the national network, there is still controversy over the next step, the creation of a commercial, ubiquitous, fiber-optic system that can reach into almost every American home. Various sectors of the communications industry—the long-distance telephone companies, the regional Bell operating companies, the cable companies—each want a central role.

It will be up to federal and state policymakers to determine how best to build a universal, high-speed network, to reconcile the competing corporate interests and to create a network that maximizes the benefits enjoyed by all Americans. Given the pace at which networking technology is changing, our efforts will demand more than just vision, leadership and commitment. Flexibility and a willingness to change the details of the plan while constantly moving forward will be required.

Building the national network will be only the first step. That done, we could focus on a telecommunications policy that would resolve the telephone-cable debate, decide what the Baby Bells can and cannot do and confront the challenge presented by other telecommunications issues. We cannot afford to remain deadlocked. The alternative is to wait until other nations show us how to take advantage of this technology—and they will. We must move first.

Common Law for the Electronic Frontier

Networked computing challenges the laws that govern information and ownership

by Anne W. Branscomb

Cyberspace," says John P. Barlow, computer activist and co-founder of the Electronic Frontier Foundation, "remains a frontier region, across which roam the few aboriginal technologists and cyberpunks who can tolerate the austerity of its savage computer interfaces, incompatible communications protocols, proprietary barricades, cultural and legal ambiguities, and general lack of useful maps or metaphors."

It looks much the same to the legal profession. As networks become less the toys of the console cowboys and more ubiquitous in the daily lives of ordinary computer users, a new breed of lawyers is trying to adapt existing laws to the electronic frontier.

The behavior of computer users in cyberspace, a term for electronic space coined by science fiction writer William Gibson, confounds the carefully honed skills of lawyers to make sense of the morass of new uses of information technology. The electronic environment of computer networks is marked by versatility, complexity, diversity and extraterritoriality. All these characteristics pose challenges to the laws that govern generating, organizing, transmitting and archiving information.

"Computer crackers" will push the limits of network security against barriers to the free flow of information.

ANNE W. BRANSCOMB is a communications lawyer at the Harvard University Program on Information Resources Policy and visiting professor at the Annenberg School of Communications of the University of Pennsylvania. As chairman of the Communications Law Division of the American Bar Association Science and Technology Section from 1980 to 1984, she edited *Toward a Law of Global Communications Networks* and organized a number of research and professional educational projects. She has served as a member of the Commission on Freedom and Equality of Access to Information and as a consultant to the National Science Foundation, the congressional Office of Technology Assessment and the State Department. Branscomb holds degrees in political science from Harvard University and the University of North Carolina and is a graduate of the George Washington University Law School. She is the author of *Who Owns Information?,* published in 1994.

Today the exchange of information in a network can defy efforts to stop distribution by the very speed with which the deed is accomplished via satellites and optical fibers. In 1988 it took only a matter of hours for the Worm, a rogue computer program, designed by Cornell University graduate student Robert T. Morris, Jr., to circulate through and disable the Internet, the network used by scientists.

The ease with which electronic impulses can be manipulated, modified and erased is hostile to a deliberate legal system that arose in an era of tangible things and relies on documentary evidence to validate transactions, incriminate miscreants and affirm contractual relations. What have been traditionally known as letters, journals, photographs, conversations, videotapes, audiotapes and books merge into a single stream of undifferentiated electronic impulses.

The complex environment and fluid messages make it difficult to determine which version of a document or electronic envelope is the draft or review or "published" copy. The diversity of inputs and outputs also makes it difficult to determine who is author, publisher, republisher, reader or archivist.

Under the 1976 revision of the Copyright Law, one must assume that any original work is protected by an unpublished copyright until published. Consequently, when precisely a work is published and under what proviso it is released are matters of considerable legal interest. Is the electronic record a copyrightable "writing"? If so, would the copyright be held by an individual or a group? Suppose some members of a group wish to extract portions of the conversation and transmit them to others. Would that constitute fair use or misuse?

Thus, the information industry, which has thrived on a convenient arrangement between vested interests of authors, publishers and libraries, is now confronting a different economic environment. Computer users have greater freedom to reach out and draw from a digital environment chunks of data in different forms. It is difficult to sort out who is entitled to compensation or royalties from which use of data. Moreover, when a distributed network has multiple participants, it will be more complicated to determine who is entitled to claim recompense for value added.

Lawyers, for whom legal jurisdiction involves the statutory reach of persons residing within certain boundaries, are further confounded by the extraterritoriality (or nonterritoriality) of cyberspace. Electronic communities abound as nodes tied together in a network, be they independent computer bulletin boards or information network provid-

Reprinted from the September 1991 issue

ers such as CompuServe, Prodigy, or the Electronic Information Exchange System (EIES).

The earliest global networks, such as the Society for Worldwide Interbank Financial Telecommunication (SWIFT), a system that transfers funds electronically, operate under very stringent rules to which member banks subscribe when entering the system. University network services are scurrying to set up their own codes of conduct in the aftermath of the Internet Worm. How these codes of conduct will mesh with local, state and national laws will prove challenging.

A dialogue held on the Whole Earth 'Lectronic Link (WELL), a computer network operated out of Marin County, Calif., by Stewart Brand of *Whole Earth Catalog* fame, suggests that "computer crackers" will push the outer limits of network security as long as any barrier exists to the free flow of information. A recent revolt of Prodigy subscribers protesting censorship of electronic messages indicates that users want more voice in determining the rules under which they will participate.

Yet what information should be free? Weather forecasts, perhaps news of impending disasters, epidemics, pests, volcanic eruptions? It is inconceivable that we should contemplate a society in which everybody must pay a meter to learn that a hurricane is on the way. But as more and more information is provided by the private rather than the public sector, the line between what is public and what is private blurs.

Moreover, the demarcation between public and private seems to change constantly. What was yesterday freely obtainable (network sporting events) comes today on a pay-as-you-go basis (on cable channels). Indeed, one obstacle to networked videotex services has been a reluctance of consumers to pay for information in electronic form when it is available in printed form at lower cost.

The privatization and commercialization of information do not sit well with computer hackers, who look on computer networks as an open, sharing society in which the skilled contribute to the welfare of the cooperative. Yet, like pioneers on the Western frontier, they are confronted by those who wish to fence in their private domains.

Many computer professionals, for example, have objected to the proprietary control of user interfaces, which programmers need to design compatible and sometimes competitive products. One such group, known as the nuPrometheus League (for the Titan who stole fire from the gods and gave it to humankind), distributed some of the source code of Apple Computer's Macintosh to a variety of computer professionals.

The distribution was apparently a protest against aggressive litigation Apple had instigated against Hewlett-Packard Company and Microsoft Corporation. Apple claimed copyright infringement of its source code on the grounds that software sold by the defending companies appropriated the "look and feel" of its user interface.

The common law of cyberspace will evolve as users seek consensual solutions to common problems.

The letter accompanying the distribution of the source code was signed "nuPrometheus League (Software Artists for Information Dissemination)" and stated: "Our objective at Apple is to distribute everything that prevents other manufacturers from creating legal copies of Macintosh. As an organization, the nuPrometheus League has no ambition beyond seeing the genius of a few Apple employees benefit the entire world, not just dissipated by Apple through litigation and ill will."

Apple management, which had once encouraged the sharing of software on floppy disks to encourage the use of its computers, promised to prosecute "to the full extent of the law." It declared that those in possession of the code were recipients of stolen property and could be prosecuted under federal laws that prohibit mailing stolen property across state lines.

But U.S. law also has traditionally recognized that much information is free-ly available for all the world to take and develop as a shared resource. Indeed, the recent Supreme Court decision of *Feist Publications* v. *Rural Telephone Service Company* confirms that the concept of shared information is alive and well.

In this case, an independent telephone company refused to license the use of its telephone directory to Feist Publications, which packaged wider-area directories. So Feist proceeded to extract the numbers, claiming they were facts that were not susceptible to copyright protection.

The Supreme Court agreed, putting to rest the labored efforts of a long line of lower courts that attempted to craft a "sweat of the brow" theory to cover the substantial investment made by data-base providers in gathering, processing, packaging and marketing their information products.

The application of intellectual property concepts to data bases may become more complex as a result of the Feist decision. Under the copyright rubric, only original expression may be copyrighted, not facts or ideas. Computers can scan pages of data, and, presumably, as long as they do not copy the exact organization or presentation or the software programs used to sort the information, they may not be infringing the copyright of the "compilation," the legal hook on which data bases now hang their protective hats.

Unless Congress or the states clarify what in a data base can be protected, information providers will have to continue to rely largely on contracts with their users. It is not entirely clear, however, whether a contract that appears momentarily on the screen prior to use is valid if the user has had no opportunity to negotiate the terms. Surely, information will not be shared with competitors unless some quid pro quo is offered or some eleemosynary motive is apparent. Giving away the fruits of intellectual labor without fair and equitable compensation is a policy not destined to survive the rigors of a marketplace economy.

At the other end of the spectrum of the debate over public and private information are the protesters who would like to control the distribution of personal information about themselves. As more computer networks come on-line, more users are becoming aware that personal data are being gathered and correlated for someone else's purposes. The regulation of transaction-generated information, such as the records from telephone calls or credit-card purchases, is a legal nightmare waiting to happen.

This issue caught the public interest through the controversy over caller I.D., or automated number identification, which allows holders of "800" numbers—or anyone subscribing to the service—to read a caller's number from a video terminal. This feature permits marketers to greet callers by their first names but, more important, to correlate their names and addresses with purchasing habits.

In fact, the telephone number is fast becoming at least as important an identifier as the social security card or the driver's license number (which are one and the same in many states). Transaction-generated information can be mixed and matched with census data, postal codes and such other publicly available information as automobile and boat registrations, birth registrations and death certificates.

This information, which has long been available to small marketers at a prohibitive cost from such large corporate information gatherers as Equifax Marketing Decision Systems, can provide precise profiles of potential buyers. But when Equifax joined forces with Lotus Development Corporation to offer an inexpensive product called Lotus Marketplace, a set of optical disks containing data on 80 million households and more than 120 million individuals, 30,000 people wrote in to have their names removed from the data base.

Yet that data base is typical of the type of information that is available when credit cards or any electronically readable media are used to make a purchase. Kmart Corporation, Sears, Roebuck and Co. and J. L. Hudson Company collect point-of-sale data from credit-card customers and either reward frequent buyers or appeal to their tastes and buying habits. One of the most aggressive of the new "relationship marketers" is Quaker Oats, which has a data base of 35 million households to whom it sends discount coupons and tracks their redemption. Citicorp has been experimenting with a data base of two million consenting customers who shop at supermarket chains across the country.

Clearly, the uproar over Lotus Marketplace, which caused it to be canceled, indicates that citizens are becoming apprehensive about the information about them that is being collected and correlated. Yet digital data are almost totally unregulated because they do not neatly fit into any of the legal boxes we have developed over the past 100 years. Traditionally the established laws governing communications apply specifically to mail, newspaper, cable television and radio broadcasting, in their separate legal pigeonholes. Signals in a digital environment do not differentiate among voice, video and data.

Already the lines between cable and broadcasting are becoming indistinct, as are the lines between what constitutes a common carrier (the telephone network) and a content provider (newspapers, journals and books). The common carriers, such as Bell operating companies, are petitioning to become information providers as a quid pro quo for installing optical fiber to the home that will facilitate the fully digital and networked environment of the future.

Ironically, the telephone companies are being besieged by demands to exercise control over the obscenity and pornography offered on the "900" paid-access calls, as anxious parents deplore salacious conversations as well as excessive charges on their telephone bills. But in late 1990, when Prodigy, an on-line service owned by IBM and Sears, admitted to controlling the content of messages posted on electronic bulletin boards, it met with just the opposite reaction from users.

The ease with which electronic impulses can be manipulated, modified and erased is hostile to a deliberate legal system.

Prodigy was accused of censorship for curtailing an on-line discussion critical of its rates and for discontinuing several controversial computer conferences involving sexual preferences. The public debate blossomed into a full-fledged airing of the nature of open discussion on electronic media.

Prodigy contended that it had the right to screen public messages posted on bulletin boards and conferences, on the basis that it was similar to the Walt Disney Productions cable channel, a family service. Many users disagreed. They likened the electronic forums in which they participated to the coffeehouses of Paris or a street corner or a private gathering in one of their homes.

Indeed, electronic bulletin boards (many operated by individuals who allow access only to their friends) have grown up in an atmosphere of open, uninhibited discussion. The WELL, for example, operates under the assumption that each person owns and is responsible for whatever he or she says.

Whether the law will recognize such individual responsibility and control in electronic media remains to be seen. Nor is it certain that the law will absolve the network provider from liability for messages it allowed to be "published" or distributed. Indeed, there is considerable danger that the law will attempt to cover all electronic media under the same rubric without recognizing the vast diversity of networks—and diverse forms of communication within the same service—that are currently developing, some with unique cultures.

Law enforcement agents are stepping up the efforts to arrest computer hackers, a term that has come to mean unscrupulous youngsters who wreak havoc with computer systems. In February 1990 members of a hacker group called the Legion of Doom were arrested and charged with breaking into the computers of BellSouth. Three of the targets later pleaded guilty to stealing a BellSouth text file for its 911 emergency telephone system.

Then, in May 1990, a two-year Secret Service operation called Sundevil culminated in 27 simultaneous searches in which 40 computer systems were seized. One indictment for publishing the information was dropped because the defendant's lawyer convinced the authorities that the material was already published. Others spent many months awaiting indictment without the return of equipment that was seized.

Sundevil precipitated an outcry from computer professionals that constitutional rights were being trampled on, and the Electronic Frontier Foundation, devoted to protecting those rights, was established. The incident itself, however, is indicative of the confusion surrounding legal concepts that have survived the test of time in the noncomputer world but may not stand up to scrutiny in the electronic context.

Property has traditionally been something that can be transferred from place to place or person to person. Theft is defined as "depriving the owner of the use thereof." In an electronic environment, although a piece of software or a component of a data base may be valuable, the intruder who accesses it without authorization is not depriving the owners or authorized users of its use, unless the data are maliciously destroyed or distorted.

Some forms of miscreant behavior, such as introducing a message bearing a virus into a network, may not be destructive. Nevertheless, they may clog the computer's memory, denying authorized users access to their data. Any unauthorized entry whatever alarms network managers because they fear some undesirable and deleterious consequence.

A "hacker" law recently enacted in the U.K. makes unauthorized entry punishable as a misdemeanor and entry with malicious intent a felony. In the U.S., several states have passed special legislation to expand their computer crime laws to cover the insertion of viruses and other rogue programs into software or network systems.

Proposed federal legislation would expand the Computer Fraud and Abuse Act of 1986 to cover "reckless disregard" for the consequences in addition to specific intent to cause damage or harm. There is also a move to enlarge the definition of venue. Such legislation would make it possible to try an alleged perpetrator in any state through which, from which or into which some part of the transmission is found to have occurred.

Another potential legal issue surrounds the ability of computers to alter and distort data and images. With computerized capability to cut and paste images, what you see may not be what actually took place. An incident that occurred in the weeks just before the election of November 6, 1990, in Massachusetts may be a harbinger of more widespread controversies. A popular Boston television journalist, Natalie Jacobson, interviewed the two contenders in the gubernatorial race. Both candidates, John Silber and William Weld, were filmed in their homes to let the viewers see the intimate sides of their characters and family lives.

During the interview, however, Silber berated the audience for denigrating those women who, like one of his daughters (as well as his wife), chose to dedicate their lives to the care of husband, home and children. The implication of his comments was that working mothers who left children in day care were selfish and uncaring.

The outburst so ruffled working women, a large voting bloc, that the Weld campaign incorporated this segment into a political advertisement. The image of Silber, however, was enlarged and slightly distorted, so that he appeared more menacing than in the original shot. In addition, taking the conversation out of context made his words sound more threatening. The ad was widely criticized, and the Silber campaign took immediate action to have it withdrawn.

The incident raises interesting political as well as legal issues about the ease with which voice, data and images can be downloaded and manipulated almost instantaneously for retransmission. Should a temporary restraining order have been available to prevent television stations from broadcasting the spot, or would that have constituted prior restraint? Should Silber have a case against Weld for libel or slander? If so, what is the appropriate penalty, recompense or sanction?

On the other hand, should the television station have a right of action against the Weld campaign or other broadcasters for copyright infringement for carrying an excerpt from a copyrighted newsmagazine? Should the television correspondent have a right of action against Weld for distorting the image or changing its context? Was the copying and rebroadcast of the segment a "fair use" of Jacobson's work product?

Cyberspace is a frontier where territorial rights are being established and electronic environments are being differentiated.

Still other relatively unexplored legal territory surrounds the way companies conduct business over networks. A recent California case, *Revlon* v. *Logisticon, Inc.*, was the computer-age equivalent of shutting off the phones for nonpayment. On October 16, 1990, Logisticon, a software firm, brought the operations of Revlon, the cosmetics firm, to a dead halt by remotely disabling the software that managed the distribution of Revlon's products from warehouse to retail stores. Logisticon justified the curtailment of service because Revlon had refused to make the next scheduled payment under its contract.

Logisticon's action, which was characterized by a Revlon vice president as "commercial terrorism," resulted in a loss of three days of activity at Revlon's main distribution centers. Revlon raced to court, accusing Logisticon of extortion, breach of contract and trespass, among other complaints. Logisticon claimed a right to disable the software because Revlon had refused to honor the next payment.

The legal issues are myriad. Who "owned" the software installed within the Revlon business offices? Was the agreement a "purchase" of software or a "license" to use software? And does it make a difference? Did Logisticon act improperly by disengaging its software so precipitously, or was it acting within its legal rights in disabling the use of its proprietary product? Should it have disclosed to Revlon the existence of the remote capability to disable the software?

Even more important than the substance of the legal issues that will arise from data networks is the question of what group will determine which laws or operating rules shall apply. Nowhere is this more clear than in the emergence of computer bulletin boards such as the WELL.

Without question, computer bulletin boards are an electronic hybrid, parts of which may be looked on either as public or private, depending on the desires of the participants. These are analogous to mail, conversations, journals, chitchat or meetings. Under normal circumstances, this electronic environment might be considered more like a street corner where one is entitled to make informal remarks to one's intimate friends.

But many bulletin boards are accessed by users intent on "publication" for the record—scientists pursuing common interests in a research project, for instance. The cooperative writing may therefore have substantial historical, political or scientific value as a publishable research paper or journal article or treatise or textbook.

A community is usually governed by its duly enacted laws, which represent the ethical values of the group. But what if the computer conference or electronic bulletin board crosses the borders of two or more states or nations? In that case, more than one legal system may apply with rules that differ within the separate legal dominions. Should the community's laws be those of the geographic or the electronic locus? If the latter, how should the laws be promulgated? Who should administer them? What sanctions should enforce them? What institutions are responsible for resolving disputes in the global information marketplace?

Cyberspace is a frontier where territorial rights are being established and electronic environments are being differentiated in much the way the Western frontier was pushed back by voyageurs, pioneers, miners and cattlemen. And the entrepreneurs are arriving with their new institutions and information technology, in much the same way as the pony express and railroads pioneered communications networks during the 19th century.

Lawmaking is a complicated process that takes place in a larger universe than the confines of legislatures and courts. Many laws are never written. Many statutory laws are never enforced. Legal systems develop from community standards and consensual observance as well as from litigation and legislative determination. So, too, will the common law of cyberspace evolve as users express their concerns and seek consensual solutions to common problems.

Quantum Cryptography

For ages, mathematicians have searched for a system that would allow two people to exchange messages in absolute secrecy. Quantum mechanics has now joined forces with cryptology to achieve a major step in that direction

by Charles H. Bennett, Gilles Brassard and Artur K. Ekert

In his classic short story "The Gold Bug," published in 1843, Edgar Allan Poe explains the rudiments of code breaking and ventures the opinion that the human mind can break any cipher that human ingenuity could devise. During the subsequent century and a half, the contest between code makers and code breakers has undergone reversals and complications that would have delighted Poe. An unbreakable cipher was invented in 1918, although its unbreakability was not proved until the 1940s. This cipher was rather impractical because it required the sender and receiver to agree beforehand on a key—a large stockpile of secret random digits, some of which were used up each time a secret message was transmitted. More practical ciphers with short, reusable keys, or no secret key at all, were developed in the 1970s, but to this day they remain in a mathematical limbo, having neither been broken nor proved secure.

A recent unexpected development is the use of quantum mechanics to perform cryptographic feats unachievable by mathematics alone. Quantum cryptographic devices typically employ individual photons of light and take advantage of Heisenberg's uncertainty principle, according to which measuring a quantum system in general disturbs it and yields incomplete information about its state before the measurement. Eavesdropping on a quantum communications channel therefore causes an unavoidable disturbance, alerting the legitimate users. Quantum cryptography exploits this effect to allow two parties who have never met and who share no secret information beforehand to communicate in absolute secrecy under the nose of an adversary. Quantum techniques also assist in the achievement of subtler cryptographic goals, important in the post–cold war world, such as enabling two mutually distrustful parties to make joint decisions based on private information, while compromising its confidentiality as little as possible.

CHARLES H. BENNETT, GILLES BRASSARD and ARTUR K. EKERT share a deep interest in the fundamental connections between physics and the theory of computation. Bennett and Brassard have pioneered the field of quantum cryptography. Since 1973 Bennett has been a researcher at the IBM Thomas J. Watson Research Center in Yorktown Heights, N.Y. In 1979 Brassard became a professor of computer science at the Université de Montréal, where he held the E.W.R. Steacie Memorial Fellowship. He is editor-in-chief of the *Journal of Cryptology*. Ekert, who received his doctoral degree from the University of Oxford, is Royal Society Howe Research Fellow at Oxford's Clarendon Laboratory.

The art of cryptography began at least 2,500 years ago and has played an important role in history ever since. Perhaps one of the most famous cryptograms, the Zimmermann Note, propelled the U.S. into World War I. When the cryptogram was broken in 1917, Americans learned that Germany had tried to entice Mexico to join its war effort by promising Mexico territories in the U.S.

Around this time Gilbert S. Vernam of American Telephone and Telegraph Company and Major Joseph O. Mauborgne of the U.S. Army Signal Corps developed the first truly unbreakable code called the Vernam cipher [*see box on page 166*]. One distinctive feature of the code is its need for a key that is as long as the message being transmitted and is never reused to send another message. (The Vernam cipher is also known as the one-time pad from the practice of furnishing the key to spies in the form of a tear-off pad, each sheet of which was to be used once and then carefully destroyed.) The discovery of the Vernam cipher did not create much of a stir at the time, probably because the cipher's unbreakability was not definitively proved until later and because its massive key requirements made it impractical for general use.

Because of this limitation, soldiers and diplomats continued to rely on weaker ciphers using shorter keys. Consequently, during World War II, the Allies were able to read most of the secret messages transmitted by the Germans and Japanese. These ciphers, though breakable, were by no means easy to crack. Indeed, the formidable task of breaking increasingly sophisticated ciphers was one of the factors that stimulated the development of electronic computers.

Academic interest in cryptology grew more intense in the mid-1970s, when Whitfield Diffie and Martin E. Hellman of Stanford University and Ralph C. Merkle of the University of California at Berkeley discovered the principle of public-key cryptography (PKC). In 1977 Ronald L. Rivest, Adi Shamir and Leonard M. Adleman, then at the Massachusetts Institute of Technology, devised a practical implementation [see "The Mathematics of Public-Key Cryptography," by Martin E. Hellman; SCIENTIFIC AMERICAN, August 1979].

Public-key cryptosystems differ from all previous schemes in that parties wishing to communicate do not need to agree on a secret key beforehand. The idea of PKC is for a user, whom we shall call Alice, to choose randomly a pair of mutually inverse transformations to be used for encryption and decryption; she then publishes the instructions for performing encryption but not decryption. Another user, Bob, can then use Alice's public-encryption algorithm to prepare a message that only she can decrypt. Similarly, anyone, including Alice, can use Bob's public-encryption algorithm to prepare a message that only he can decrypt. Thus, Alice and Bob can converse secretly even though they share no secret to begin with. Public-key cryptosystems are especially suitable for encrypting electronic mail and commercial transactions, which often occur between parties who, unlike diplomats and spies, have not anticipated their need to communicate secretly.

Offsetting this advantage is the fact

Reprinted from the October 1992 issue

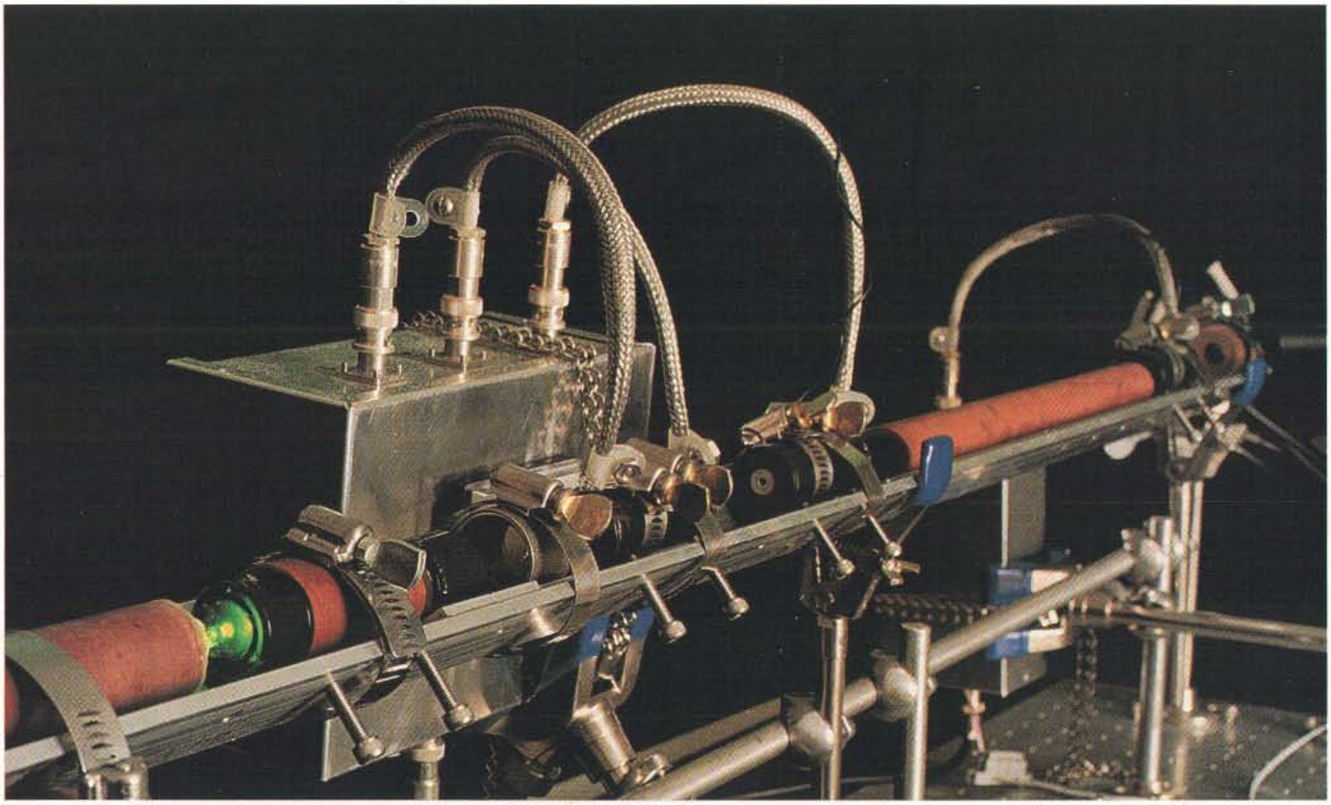
that public-key systems have not been proven to be secure. Indeed, in 1982 Shamir, now at the Weizmann Institute of Science in Israel, cracked one of the early public-key cryptosystems, the knapsack cipher. Poe could be smiling from the grave, knowing there is a clever method of attack, as yet undiscovered, that could break any of these ciphers in a few minutes.

Several years before the discovery of public-key cryptography, another striking development had quietly taken place: the union of cryptography with quantum mechanics. Around 1970 Stephen J. Wiesner, then at Columbia University, wrote a paper entitled "Conjugate Coding," explaining how quantum physics could be used, at least in principle, to accomplish two tasks that were impossible from the perspective of classical physics. One task was a way to produce banknotes that would be physically impossible to counterfeit. The other was a scheme for combining two classical messages into a single quantum transmission from which the receiver could extract either message but not both. Unfortunately, Wiesner's paper was rejected by the journal to which he sent it, and it went unpublished until 1983. Meanwhile, in 1979, two of us (Bennett and Brassard) who knew of Wiesner's ideas began thinking about how to combine them with public-key cryptography. We soon realized that they could be used as a substitute for PKC: two users, who shared no secret initially, could communicate secretly, but now with absolute and provable security, barring violations of accepted physical laws.

Our early quantum cryptographic schemes, developed between 1982 and 1984, were somewhat impractical, but refinements over the next few years culminated in the building of a fully working prototype at the IBM Thomas J. Watson Research Center in 1989. John Smolin, now at the University of California at Los Angeles, helped to build the electronics and optics for the apparatus, and François Bessette and Louis Salvail of the University of Montreal assisted in writing the software. At about the same time, the theoretical ideas of David Deutsch of the University of Oxford led one of us (Ekert) to conceive of a slightly different cryptosystem based on quantum correlations. In early 1991, utilizing ideas conceived by Massimo Palma of the University of Palermo, John Rarity and Paul Tapster of the British Defence Research Agency started experiments implementing Ekert's cryptosystem.

To explain how such systems work, we need to describe in more detail some aspects of the mathematics of classical cryptography, especially the role of the key. In the early days of cryptography the security of a cipher depended on the secrecy of the entire encryption and decryption procedure. Today such procedures are usually known publicly, but the key is kept secret. In such ciphers the key is used to control and customize the encryption and decryption processes in such a way that an adversary who has intercepted the cryptogram and knows the general method of encryption but not the key will not be able to infer anything useful about the original message. Consequently, the cryptogram may be broadcast over a public channel such as a radio or printed in a newspaper. The key, however, must be sent through a very secure private channel, such as a clandestine meeting or a delivery by a trusted courier. Although the distribution of a key over private channels is expensive, it makes possible subsequent secret communication over inexpensive public channels.

Ultimately, the security of a cryptogram depends on the length of the key. In two brilliant papers written in the 1940s, Claude E. Shannon, then at Bell Laboratories, showed that if the key is shorter than the message being encrypted, some information about the message can be inferred from the cryptogram by a sufficiently powerful adversary. This leakage of information occurs



QUANTUM DEVICE generates and measures extremely faint flashes of polarized light, providing a secure way to transmit information [*see illustration on pages 170 and 171*]. On average, each flash consists of one tenth of a photon.

regardless of how complicated the encryption process may be. In contrast, the message can be completely and unconditionally hidden from the eavesdropper by cryptosystems such as the Vernam cipher, in which the key is as long as the message, is purely random and is used only once.

Even the Vernam cipher is only as secure as its key distribution and key storage. Because of the great difficulty of supplying secure new keys for every message, the Vernam cipher is impractical for general commercial uses, although it is routinely employed for diplomatic communications such as those exchanged over the Moscow-Washington hot line. In contrast, the most widely used commercial cipher, the Data Encryption Standard, depends on a 56-bit secret key, which is reused for many encryptions over a certain length of time. This scheme simplifies the problem of secure key distribution and storage, but it does not eliminate it.

A fundamental problem remains. In principle, any classical private channel can be monitored passively, without the sender or receiver knowing that the eavesdropping has taken place. For example, a key carried by a trusted courier might have been read en route by a surreptitious high-resolution x-ray scan or other sophisticated imaging technique without the courier's knowledge. More generally, classical physics—the theory of macroscopic bodies and phenomena such as paper documents, magnetic tapes and radio signals—allows all physical properties of an object to be measured without disturbing those properties. Since all information, including a cryptographic key, is encoded in measurable physical properties of some object or signal, classical theory leaves open the possibility of passive eavesdropping, because it allows the eavesdropper in principle to measure physical properties without disturbing them.

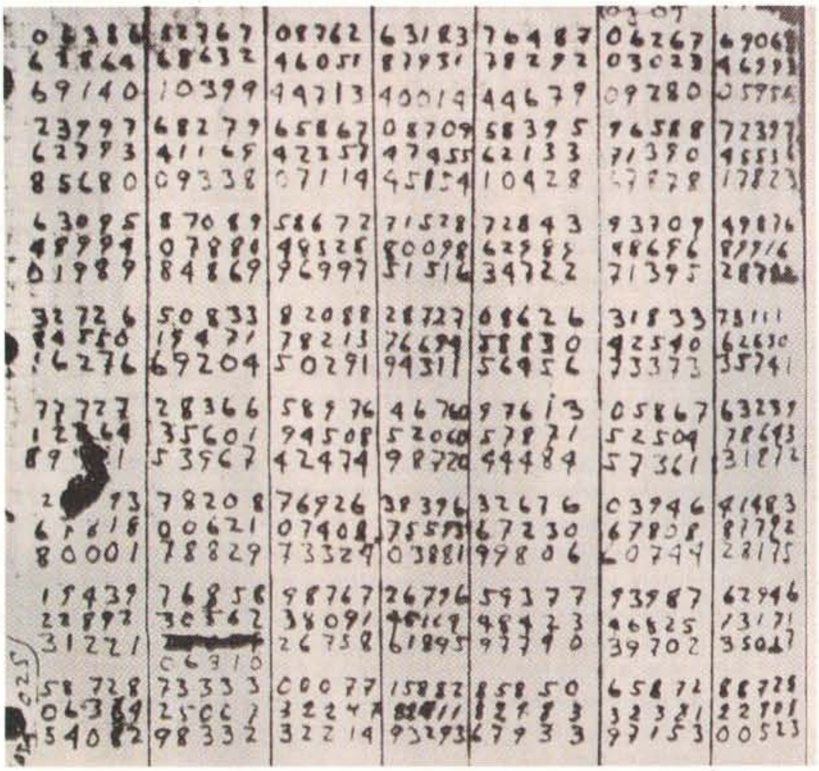
This is not the case in quantum theory, which forms the basis for quantum cryptography. Quantum theory is believed to govern all objects, large and small, but its consequences are most conspicuous in microscopic systems such as individual atoms or subatomic particles. The act of measurement is an integral part of quantum mechanics, not just a passive, external process as in classical physics. So it is possible to design a quantum channel—one that carries signals based on quantum phenomena—in such a way that any effort to monitor the channel necessarily disturbs the signal in some detectable way. The effect arises because in quantum

The Cipher of Ché Guevara

When in 1967 the Bolivian army captured and executed the revolutionary Ché Guevara, they found on his body a worksheet showing how he prepared a message for transmission to Cuban president Fidel Castro. Guevara used the unbreakable cipher invented by Gilbert Vernam in 1918. The letters of Guevara's message (in Spanish) were first translated into one- and two-digit decimal numbers by a fixed rule, namely:

A 6	E 8	I 39	M 70	Q 71	U 52	Y 1
B 38	F 30	J 31	N 76	R 58	V 50	Z 59
C 32	G 36	K 78	O 9	S 2	W 56	
D 4	H 34	L 72	P 79	T 0	X 54	

By itself this procedure would have provided virtually no protection. The message digits were then strung together in convenient five-digit blocks. They became the top line of each three-line group on the worksheet. The middle line of each group is the key, a sequence of random digits known only to Guevara and Castro.



Next the message and key were added (without carries) to produce a cryptogram, forming the bottom line of each three-line group. Because of the addition of the random key digits, this cryptogram is itself a random decimal sequence, carrying no information about the original message, except to someone who knows the key. The cryptogram was then transmitted to Cuba by an insecure channel such as shortwave radio. At the receiving end, Castro's cipher office would have subtracted the same random key digits, reconstructing the number sequence in the top row, and then would have translated the numbers back into the letters of the message.

Many spies and diplomats have used the Vernam cipher throughout the 20th century. The key, rather than consisting of decimal digits, can be a long random sequence of the binary digits 0 and 1, and the additions and subtractions would be carried out in base 2 by machine, rather than in base 10 by hand. Nevertheless, the key must still be hand-carried from the place where it is generated to the places where it will be used, and it must be assiduously guarded during all phases of delivery and storage to prevent it from falling into the hands of an adversary.

theory, certain pairs of physical properties are complementary in the sense that measuring one property necessarily disturbs the other. This statement, known as Heisenberg's uncertainty principle, does not refer merely to the limitations of a particular measurement technology: it holds for all possible measurements.

The uncertainty principle can be applied to design a completely secure channel based on the quantum properties of light. The smallest unit, or quantum, of light is the photon, which can be thought of as a tiny, oscillating electric field. The direction of the oscillation is known as the photon's polarization. Ordinary light consists of photons that have many different polarizations. But if the light passes through a polarizing filter, such as those used in sunglasses, only photons having a particular polarization will make it through. Which polarization is transmitted depends on the orientation of the filter. In sunglasses the filters are oriented to transmit vertically polarized light because such light reflects off most horizontal surfaces with less glare. But if the glasses are turned 90 degrees, so that one lens is directly above the other, they will transmit horizontally polarized light, augmenting the glare instead of diminishing it.

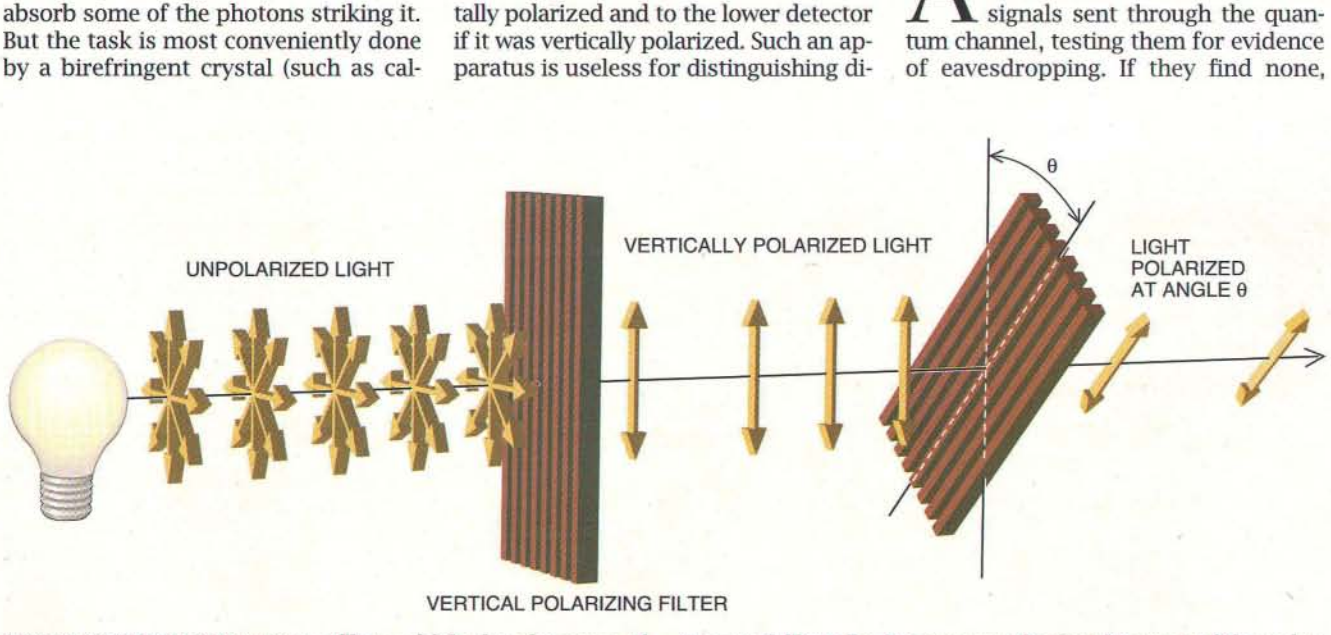
To construct a quantum channel, one needs a polarizing filter or other means for the sender to prepare photons of selected polarizations and a way for the receiver to measure the polarization of the photons. The latter job could be accomplished by another polarizing filter, which would absorb some of the photons striking it. But the task is most conveniently done by a birefringent crystal (such as calcite), which sends incident photons, depending on their polarization, into one of two paths without absorbing any [*see illustration on next page*].

A photon encountering a calcite crystal behaves in one of two ways depending on its polarization in relation to the crystal. The photon may pass straight through the crystal and emerge polarized perpendicular to the crystal's optic axis, or it may be shifted and emerge polarized along that axis. If the photon entering the crystal is already polarized in one of these two directions, it will undergo no change of polarization but will be deterministically routed into the straight or shifted path, respectively. If a photon polarized at some intermediate direction enters the crystal, however, it will have some probability of going into each beam and will be repolarized according to which beam it goes into, forgetting its original polarization. The most random behavior occurs when the photon is polarized halfway between these two directions, that is, at 45 or 135 degrees. Such photons are equally likely to go into either beam, revealing nothing about their original polarization and losing all memory of it.

Suppose Bob is told in advance that a given photon is polarized in one of the two "rectilinear" directions, vertical (90 degrees) or horizontal (0 degrees) without being informed of the specific polarization. Then he can reliably tell which direction by sending the photon into an apparatus consisting of a vertically oriented calcite crystal and two detectors, such as photomultiplier tubes, that can record single photons. The calcite crystal directs the incoming photon to the upper detector if it was horizontally polarized and to the lower detector if it was vertically polarized. Such an apparatus is useless for distinguishing diagonal (45- or 135-degree) photons, but these can be reliably distinguished by a similar apparatus that has been rotated 45 degrees from the original orientation. The rotated apparatus, in turn, is useless for distinguishing vertical from horizontal photons. According to the uncertainty principle, these limitations apply not just to the particular measuring apparatus described here but to any measuring device whatsoever. Rectilinear and diagonal polarizations are complementary properties in the sense that measuring either property necessarily randomizes the other.

We can now describe the simple scheme for quantum key distribution that two of us (Bennett and Brassard) proposed in 1984 and that we dubbed "BB84." The purpose of the scheme is for Alice and Bob to exchange a secret random key that they can subsequently use, as in the Vernam cipher, to send meaningful secret messages when the need arises. Like other quantum key distribution schemes, BB84 uses a quantum channel, through which Alice and Bob send polarized photons, in conjunction with a classical public channel, through which they send ordinary messages. An eavesdropper, whom we shall call Eve, is free to try to measure the photons in the quantum channel, but, as noted above, she cannot in general do this without disturbing them. Furthermore, she learns the entire contents of messages sent through the public channel, but assume for the moment that she could not disturb or alter these messages even if she wanted to.

Alice and Bob use the public channel to discuss and compare the signals sent through the quantum channel, testing them for evidence of eavesdropping. If they find none,

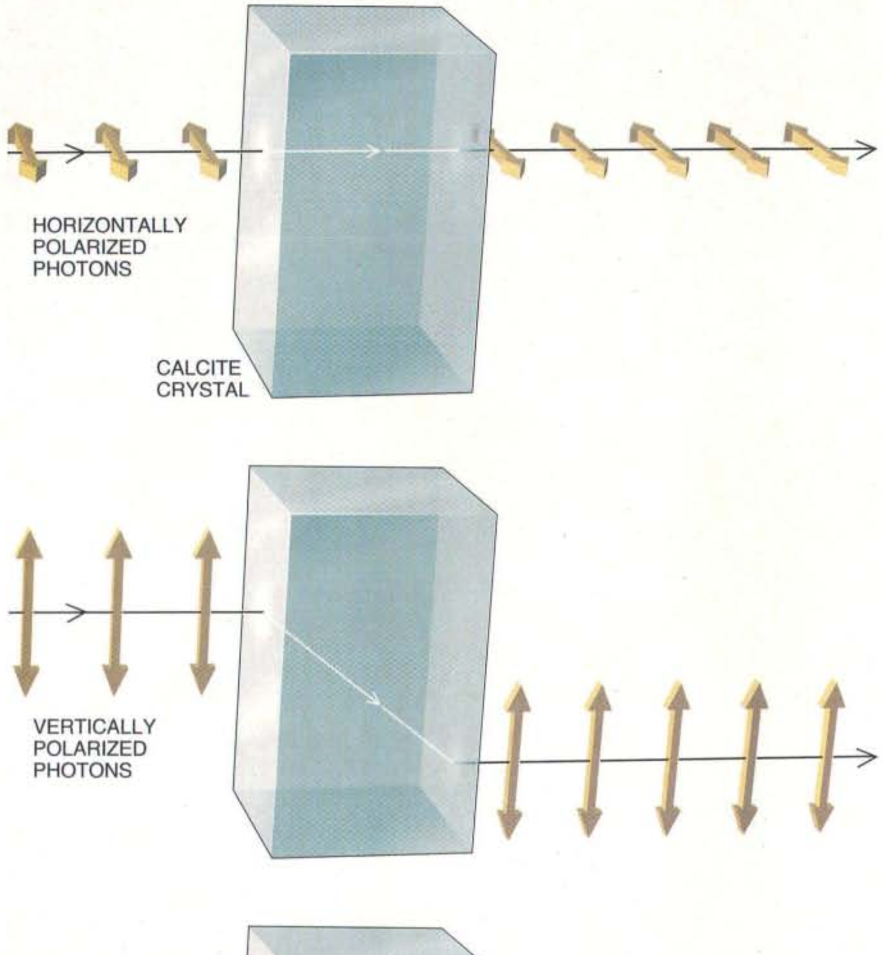


UNPOLARIZED LIGHT enters a filter, which absorbs some of the light and polarizes the remainder in the vertical direction. **A second filter tilted at some angle absorbs some of the polarized light and transmits the rest, giving it a new polarization.**

they can distill from their data a body of information that is certifiably shared, random and secret, regardless of Eve's technical sophistication and the computing power at her disposal [*see box on opposite page*]. The scheme works as follows:

First, Alice generates and sends Bob a sequence of photons whose polarizations she has chosen at random to be either 0, 45, 90 or 135 degrees. Bob receives the photons and, for each photon, decides randomly whether to measure its rectilinear or diagonal polarization.

Next Bob announces publicly, for each photon, which type of measurement he has made (rectilinear or diagonal) but not the measurement result (for example, 0, 45, 90 or 135 degrees). Alice tells him publicly, for each photon, whether he has made the right kind of measurement. Alice and Bob then discard all cases in which Bob has made the wrong measurement or in which his detectors have failed to register a photon at all (existing detectors are not 100 percent efficient). If no one has eavesdropped on the quantum channel, the remaining polarizations should be shared secret information between Alice and Bob.



CALCITE CRYSTAL can be used to distinguish between horizontally and vertically polarized photons. Horizontally polarized photons pass straight through, whereas vertically polarized photons are deflected. When diagonally polarized photons enter the crystal, they are repolarized at random in either the vertical or horizontal direction and are shifted accordingly.

Alice and Bob next test for eavesdropping, for example, by publicly comparing and discarding a randomly selected subset of their polarization data. If the comparison test shows evidence of eavesdropping, Alice and Bob discard all their data and start over with a fresh batch of photons. Otherwise they adopt the remaining polarizations, which have never been publicly mentioned, as shared secret bits, interpreting horizontal or 45-degree photons as binary 0's and vertical or 135-degree photons as binary 1's.

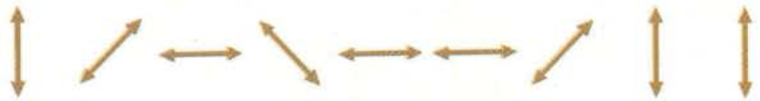
Because of the uncertainty principle, Eve cannot measure both rectilinear and diagonal polarizations of the same photon. If, for a particular photon, she makes the wrong measurement, then, even if she resends Bob a photon consistent with the result of her measurement, she will have irretrievably randomized the polarization originally sent by Alice. The net effect is to cause errors in one quarter of the bits in Bob's data that have been subjected to eavesdropping.

Directly comparing selected bits to test for errors, as described above, is not very efficient: too many bits must be sacrificed in order to build reasonable confidence that Alice's and Bob's data are identical, especially if the eavesdropping has been rather infrequent, resulting in only a few errors. A much better idea is for Alice and Bob to compare the "parity"—evenness or oddness—of a publicly agreed-on random subset containing about half the bits in their data. For instance, Alice could tell Bob, "I looked at the 1st, 3rd, 4th, 9th,...996th and 999th of my 1,000 bits of data, and they include an even number of 1's." Bob would then count the number of 1's that he has in those same locations. If he finds an odd number of 1's, he can immediately conclude that his data are different from Alice's. It can be shown that if Alice's and Bob's data differ, comparing the parity of a random subset will detect that fact with a probability of $\frac{1}{2}$, regardless of the number and location of errors. It suffices to repeat the test 20 times, with 20 different random subsets, to reduce the chance of an undetected error to less than one in a million.

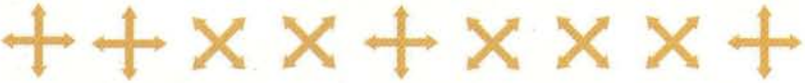
The BB84 scheme was modified to produce a working quantum cryptography apparatus at IBM [*see illustration on pages 170 and 171*]. The modifications were necessary to deal with practical problems such as noise in the detectors and the fact that the prototype uses dim flashes of light instead of single photons.

Quantum Key Distribution

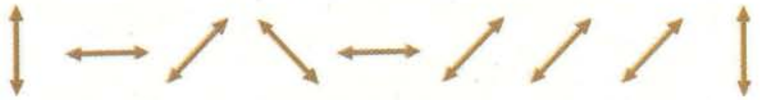
A quantum cryptographic system will allow two people, say, Alice and Bob, to exchange a secret key. The system includes a transmitter and a receiver. Alice uses the transmitter to send photons in one of four polarizations: 0, 45, 90 or 135 degrees. Bob uses the receiver to measure the polarization. According to the laws of quantum mechanics, the receiver can distinguish between rectilinear polarizations (0 and 90), or it can quickly be reconfigured to discriminate between diagonal polarizations (45 and 135); it can never, however, distinguish both types. The key distribution requires several steps. Alice sends photons with one of four polarizations, which she has chosen at random.



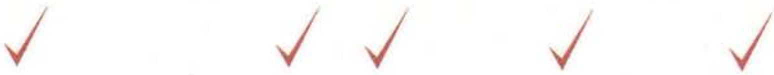
For each photon, Bob chooses at random the type of measurement: either the rectilinear type (+) or the diagonal type (×).



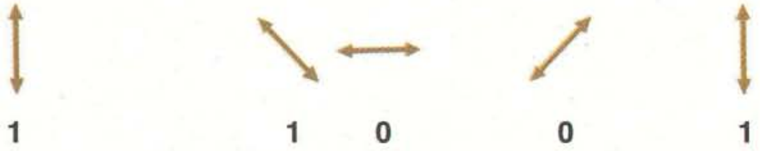
Bob records the result of his measurement but keeps it a secret.



Bob publicly announces the type of measurements he made, and Alice tells him which measurements were of the correct type.



Alice and Bob keep all cases in which Bob measured the correct type. These cases are then translated into bits (1's and 0's) and thereby become the key.



The quantum channel, with Alice's sending apparatus at one end and Bob's receiving apparatus at the other, is housed in a light-tight box. During operation, the system is controlled by a personal computer, which contains separate software representations of Alice, Bob and, optionally, Eve.

The leftmost part of Alice's sending apparatus consists of a green light-emitting diode, a pinhole, a lens and filters that provide a collimated beam of horizontally polarized light. Next, electro-optic devices known as Pockels cells are used to change the original horizontal polarization to any of four standard polarization states under Alice's control. This has the same effect as mechanically rotating her polarizing filter into four different positions, but the process can be done much faster.

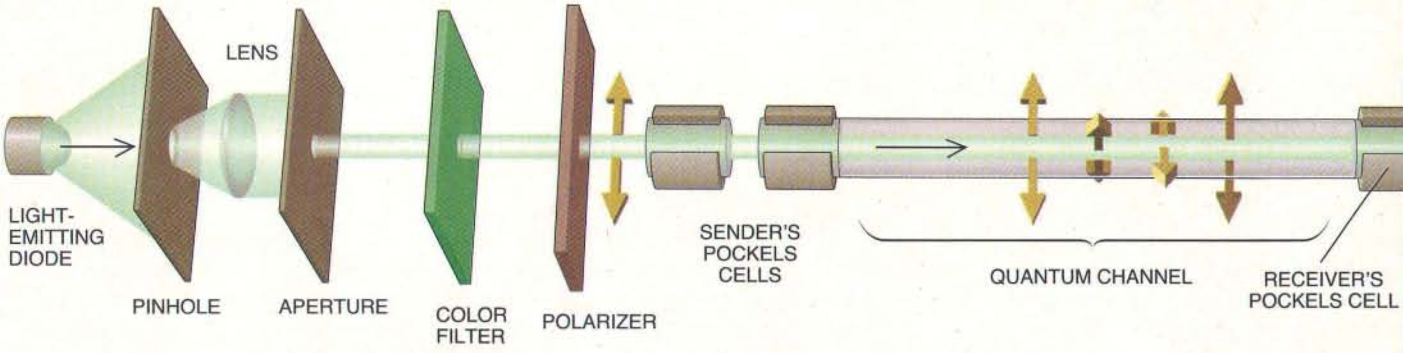
At the other end of the quantum channel, Bob's receiving apparatus contains a similar Pockels cell, which allows him to choose the type of polarization he will measure, again without actually rotating his detecting device. After the beam passes through Bob's Pockels cell, it is split by a calcite prism into two perpendicularly polarized beams, which are directed into two photomultiplier tubes for the purpose of detecting individual photons.

The sending and receiving apparatuses in the prototype are only about 30 centimeters apart—chiefly because of the desire to keep the device of reasonable dimensions for a desktop—but nothing in principle prevents the technique from being used over much greater distances. For example, quantum transmissions could be sent through several kilometers of optical fiber. If cost and inconvenience were no concern, quantum transmissions could be sent over arbitrarily great distances with negligible losses through an evacuated straight pipe. But quantum key distribution has to compete with classical techniques, which are much cheaper over long distances and may be sufficiently secure.

Recall that the BB84 scheme encodes each bit in a single polarized photon. In contrast, the prototype encodes each bit in a dim flash of light. This introduces a new eavesdropping threat to the system: if Eve taps into the beam by a device such as a half-silvered mirror, she will be able to split each flash into two flashes of lesser intensity, reading one herself while letting the other pass to Bob, its polarization undisturbed. If Eve diverts only a modest fraction of the beam, Bob may not notice the weakening of his signal, or he may attribute it to natural losses in the channel. This attack can be effectively thwarted, at the cost of reducing the rate of data transmission through the quantum channel, by having Alice send very dim flashes—that is, an intensity of less than one photon per flash on average. Such extremely dim flashes can easily be made by filtering out almost all the intensity of a bright flash.

When low-intensity flashes are used, Bob's chance of detecting a photon in any given flash is of course proportionately reduced, but the chance of Bob and Eve both detecting photons in the same flash is reduced even more, as the square of the intensity. The actual apparatus generates intensities of about one tenth of a photon per flash. On the other hand, if Alice's flashes were much brighter (say, thousands of photons per flash), they would be easy prey for the beam-splitting attack. By splitting off only a small fraction of the intensity, Eve could still obtain enough photons from each flash to perform both rectilinear and diagonal measurements and so determine the polarization unambiguously. In other words, the brighter Alice's flashes, the more they behave like classical signals, about which an eavesdropper can gain complete information while introducing negligible disturbance.

Another practical problem arises from the fact that available detectors sometimes produce a response even when no photon has arrived. Such "dark



QUANTUM SYSTEM can distribute information in perfect secrecy. The transmitter produces faint flashes of green light from a light-emitting diode. The pinhole, lens and filter create a collimated beam of dim flashes. The light is then polarized horizontally. Two Pockels cells change the polarization to 0, 45, 90 or 135 degrees. The polarized light flashes are released from the transmitter and eventually reach the receiver. There another Pockels cell shifts the polarization by either 45 degrees or not at all. The action of this Pockels cell allows the receiver to choose between measuring rectilinear or diagonal

counts" and other imperfections in the apparatus lead to errors even when there has been no eavesdropping and make it impractical for Alice and Bob simply to reject their data whenever they find an error in them, as is done in the ideal BB84 protocol. If Bob and Alice find a small number of errors, they must devise a way to correct them and proceed. On the other hand, if they find a large number, indicating significant eavesdropping, they must reject their data and start over.

A variety of techniques are available for Alice and Bob to correct a small number of errors through public discussion, such as the use of error-correcting codes. But these techniques potentially leak information to Eve, who may be listening to the public discussion. Therefore, after the quantum transmission and the error-correcting discussion, Alice and Bob find themselves with what might be thought of as an impure key, a shared body of data that is only partly secret. Information on that key may have leaked to Eve at several stages. She may have gained information by splitting some flashes, by directly measuring others (she cannot do this too often, as it causes errors in Bob's data) and by listening to the public discussion between Alice and Bob. Fortunately, Alice and Bob, because they know the intensity of the light flashes and the number of errors found and corrected, can estimate how much information might have leaked to Eve through all these routes.

In itself, such an impure key is almost worthless. If it were used as a key for the Vernam cipher, for example, it might prove very insecure if the most important part of the message happened to coincide with a part of the key the eavesdropper knew. Fortunately, two of us (Bennett and Brassard), in collaboration with Jean-Marc Robert (then a student of Brassard), developed a mathematical technique known as privacy amplification. Using this technique, Alice and Bob, through public discussion, can take such a partly secret key and distill from it a smaller amount of highly secret key, of which the eavesdropper is very unlikely to know even one bit. The essential idea of privacy amplification is for Alice and Bob, after the eavesdropping has taken place, to choose publicly a length-reducing transformation to apply to their impure key so that partial information about the input conveys almost no knowledge of the output.

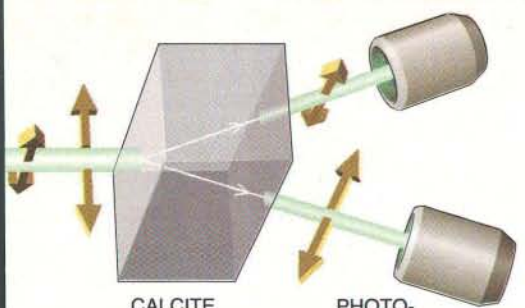
For example, if the input consists of 1,000 bits about which Eve knows at most 200 bits, Alice and Bob can distill nearly 800 highly secret bits as output. Fairly simple techniques can be shown to suffice, and Alice and Bob do not even need to know which partial information the eavesdropper might have about the input in order to choose a transformation about whose output Eve has almost no information. In particular, it suffices for Alice and Bob to define each bit of the output as the parity of an independent, publicly agreed-on random subset of the input bits, very much as they had done to gain high confidence that their raw quantum data were identical (except that now they keep the parity secret instead of publicly comparing it).

The problem of key security is not entirely solved by secure key distribution. Another weak point is key storage. Once Alice and Bob have established the key, they must store it until it is needed. But the longer they keep the key in, say, their secret safe, the more vulnerable it is to unauthorized inspection. Although principles of engineering may make the safe difficult to crack, the laws of physics always provide the possibility of a security breach. Surprisingly, a cryptosystem can be designed that can guarantee the security of both key distribution and storage by employing quantum correlations. The cryptosystem is based on David Bohm's version of the famous Einstein-Podolsky-Rosen (EPR) effect [see "The Quantum Theory and Reality," by Bernard d'Espagnat; SCIENTIFIC AMERICAN, November 1979].

The EPR effect occurs when a spherically symmetric atom emits two photons in opposite directions toward two observers, Alice and Bob. The two photons are produced in an initial state of undefined polarization. But because of the symmetry of the initial state, the polarizations of the photons, when measured, must have opposite values, provided that the measurements are of the same type. For example, if Alice and Bob both measure rectilinear polarizations, they are each equally likely to record either a 0 (horizontal polarization) or a 1 (vertical), but if Alice obtains a 0, Bob will certainly obtain a 1, and vice versa.

The unusual and important aspect of the EPR effect is that the polarization of both photons is determined as soon as, but not before, one of the photons is measured. This happens no matter how far apart the photons may be at the time. This "classical" explanation of the EPR effect is somewhat counterintuitive, and indeed all classical explanations of the EPR effect involve some implausible element, such as instantaneous action at a distance. Yet the mathematical formalism of quantum mechanics accounts for the EPR effect in a straightforward manner, and experiments have amply confirmed the existence of the phenomenon.

Employing the EPR effect, one of us (Ekert) recently devised a cryptosystem that guarantees the security of both



polarization. In the rectilinear case, a horizontally polarized photon will be directed toward the right photomultiplier; a vertically polarized photon will be directed toward the left photomultiplier.

key distribution and key storage. In a simplified version of this system, as described by N. David Mermin of Cornell University, Alice generates a number of EPR photon pairs, keeping one member of each pair for herself while sending the other to Bob. Alice and Bob measure some of their photons immediately to test for eavesdropping but store the remainder without measuring them. Then, just before the key is needed, they measure and compare some of the stored photons. If no one has tampered with the photons in storage, Bob will always obtain 1 when Alice obtains 0, and vice versa. If no discrepancies are found, Alice and Bob measure the remaining stored photons to obtain the desired key.

Although this procedure works in theory, it cannot be used in practice because of the technical infeasibility of storing photons for more than a small fraction of a second. Thus, at present, the EPR effect is not a practical means for certifying the security of key storage.

Although the best-known application of cryptography is secret communication, two other applications are probably more important in peacetime. The first is authentication: certifying that a message is from whom it says it is from and has not been altered in transit. The second is the task of maintaining the confidentiality of private information while it is being used to reach public decisions.

For most of recorded history, authentication has depended on physical objects that are hard to copy, such as seals and signatures. Such devices provide limited security, and they cannot be used at all for digital electronic documents, such as bank transactions, which are often transmitted over insecure telecommunications lines.

Fortunately, several mathematical techniques are available for authenticating digital messages. In 1979 Mark N. Wegman and J. Lawrence Carter of IBM discovered a digital authentication scheme that does provide provable security. Like the Vernam cipher, it requires that the sender and receiver possess beforehand a shared secret key, part of which is used up each time a message is authenticated.

Wegman-Carter authentication and quantum key distribution can each benefit the other. On the one hand, the quantum technique supplies the secret key bits consumed by the authentication method. On the other, Wegman-Carter authentication can be used to conduct quantum key distribution successfully even in the presence of a more powerful adversary, that is, someone who could alter the public channel messages as well as listening to them.

Quantum cryptography may also prove useful for protecting private information while it is being used to make public decisions. The classic example of such discreet decision making is the "dating problem," in which two singles seek a way of making a date if and only if each likes the other, without disclosing any further information. For example, if Alice likes Bob, but Bob does not like Alice, the date should be called off without Bob's finding out that Alice likes him. (On the other hand, it is logically unavoidable for Alice to learn that Bob does not like her, because if he did the date would be on.)

There are many other situations in which joint decision making between corporate or governmental organizations or between an individual and an organization depends on confidential information that the negotiating parties would rather not disclose in full [see "Achieving Electronic Privacy," by David Chaum; SCIENTIFIC AMERICAN, August 1992]. A discreet solution to the dating problem, or any other joint decision depending on private data, can be obtained by having Alice and Bob confide their private data to a trusted intermediary (Eve, for example) and letting her make the decision. The hazards of this approach are obvious: Alice and Bob must trust Eve both to make the decision correctly in the first place and never to reveal the private data.

Various other techniques allow public decisions based on private inputs to be made without a single trusted intermediary. For example, a successful protocol can be set up among many participants that will fail only if a majority of the participants conspire to spoil the output or reveal the inputs. On the other hand, if two parties believe in the security of public-key cryptosystems, they can make decisions discreetly without any intermediary. In 1982 Andrew C.-C. Yao, then at Stanford University, was one of the first to investigate this problem.

Recently Claude Crépeau of the École Normale Supérieure and CNRS and his student Marie-Hélène Skubiszewska, in collaboration with two of us (Bennett and Brassard), have shown that a quantum apparatus similar to the one already built for key distribution can be used to make joint decisions discreetly without intermediaries and without unproved mathematical assumptions. Discreet decision making can be implemented by repeated application of a curious information-processing procedure called oblivious transfer. The procedure is a version of Wiesner's feat of sending two messages in such a way that the receiver can read either one but not both. In 1981 Michael O. Rabin of Harvard University formalized the notion of oblivious transfer, unaware of Wiesner's pioneering but unpublished work a decade earlier. Later, Crépeau, Joe Kilian, then at the Massachusetts Institute of Technology, and others demonstrated that oblivious transfer could be applied to discreet decision making.

One particularly attractive feature of discreet quantum decision making is that, unlike key distribution, it is worthwhile even over short distances. Unfortunately, known implementations are mathematically rather inefficient, requiring many thousands of photons to be sent and received to reach even simple decisions. If its mathematical efficiency can be sufficiently improved, discreet decision making may become the most important application of quantum cryptography.

FURTHER READING

THE CODEBREAKERS: THE STORY OF SECRET WRITING. David Kahn. Macmillan, 1967.

MODERN CRYPTOLOGY: A TUTORIAL. Gilles Brassard in *Lecture Notes in Computer Science,* Vol. 325. Springer-Verlag, 1988.

QUANTUM CRYPTOGRAPHY BASED ON BELL'S THEOREM. Artur K. Ekert in *Physical Review Letters,* Vol. 67, No. 6, pages 661-663; August 5, 1991.

EXPERIMENTAL QUANTUM CRYPTOGRAPHY. Charles H. Bennett, François Bessette, Gilles Brassard, Louis Salvail and John Smolin in *Journal of Cryptology,* Vol. 5, No. 1, pages 3-28; 1992.

QUANTUM CRYPTOGRAPHY WITHOUT BELL'S THEOREM. Charles H. Bennett, Gilles Brassard and N. David Mermin in *Physical Review Letters,* Vol. 68, No. 5, pages 557-559; February 3, 1992.

Dr. Big Brother

Will health care reform make everyone's life an open book?

The Clinton administration's health plan should give every American access to medical care and protection from financial ruin. It would also eventually give the government some of the most intimate details of a person's private life. Without carefully crafted safeguards, the president's health security plan could become an instrument of Big Brother.

If the Clinton plan becomes law, a powerful seven-member National Health Board will sit down to work out the fine details for a medical information highway. This electronic health data network would over time enable regional health alliances, insurers and federal, state and local health officials and researchers to assemble and exchange information about each citizen covered (the plan will ultimately include everyone). Every American will also carry a plastic card with a patient identification number, which some privacy groups fear could turn into a de facto national identification card.

Certainly, a health information network offers clear, even compelling, benefits. Through it clinicians could quickly retrieve information about patients from anywhere in the country. Redundant medical tests could thus be avoided. Health claims could be speeded. Corrected information could simultaneously be entered in many files dispersed throughout the network. Epidemiological studies could be performed rapidly.

Such ready access also means the information could more easily be abused. Whether authorized or not, the peruser of a data file could discover details of a patient's life history that go well beyond a list of past vaccinations. The records might include information about genetic testing, sexual orientation, drug use, sexually transmitted diseases, finances and other personal details.

"These data are important because health care information can influence decisions about a person's access to credit, admission to educational institutions and his or her ability to secure employment and obtain insurance," said Paula J. Bruening, an Office of Technology Assessment policy analyst, at a congressional hearing in November 1993. "Inaccuracies in the information, or its improper disclosure," Bruening claimed, "can deny an individual access to these basic necessities of life, and can threaten their personal and financial well-being."

The Clinton administration is well aware of the threat to privacy that easy access to medical information poses. A provision of the health proposal—Title V of the Health Security Act—specifies that electronic records will not be available to employers or to others who are not directly involved with health care. Furthermore, during the program's early days the system would offer access only to insurance information, not patients' complete health records. Later, though, at least some clinical and administrative data would be added. The law also has teeth: civil penalties for unapproved disclosure of confidential information and criminal sanctions for misuse of the health card or a patient identification number.

Privacy advocates have applauded the broad outlines of the Clinton plan. Yet in the next breath, they worry about what has been left unwritten. Some protections would be in place at the time each person receives coverage under the new law, perhaps during 1996. But the National Health Board that will administer the plan would not then have to produce recommendations for comprehensive legislation concerning health care privacy until 1997. "The Clinton Act is a framework. It leaves everything up to the National Health Board," says Aimee Berenson, legislative counsel for the AIDS Action Council, an organization that has worked on protecting privacy of the medical records of AIDS patients. And, Bruening adds, "It's important to build safeguards and security at the beginning. It's more effective and less costly."

If a revamped health system relies on the Social Security number to identify each patient, the risk to personal privacy could be greatly amplified. To be sure, broadening the uses of these existing numbers would forestall the need to devise new personal identification codes. But an individual's Social Security number fails to meet the Clinton health plan's goal of giving every citizen a unique and private identifier.

Indeed, critics argue, employing the Social Security number would make it that much easier to build a composite profile of how a person lives. Health information could be added to other data made accessible with the Social Security number. The number already identifies individuals in the records of many institutions, including utilities, collection agencies, credit-card issuers, stock brokerages and the Internal Revenue Service. "Why build what should be a tamper-resistant system on a system that is broken?" comments Janlori Goldman, head of the American Civil Liberties Union's Privacy and Technology Project.

"The issue is not what number do I use. The issue is what are the allowable linkages between different data sets using that number?" replies Joan Turek-Brezina, who chairs a Department of Health and Human Services task force that examines privacy issues surrounding health records. "People know the Social Security number, and it is the cheapest and fastest way to implement all of this."

The debate will become more acute as advances in technology further complicate the trade-off between privacy and efficiency. Clinton administration officials have said they will forgo the additional expense of fitting every American's wallet with a "smart" card, equipped with an electronic memory and microprocessor, or optical storage cards, which retain large amounts of data on a medium similar to that used for recording music on compact discs. In the administration's plan, a simple magnetic stripe on the back of each card can be swiped through a card reader. The health identification number encoded on the card will validate insurance coverage. That decision will avoid the inevitable debate over the privacy implications of storing the equivalent of hundreds of pages of text on a credit-card-size device.

Officials have not, however, ruled out the smart card as future health care reforms are introduced. The prospect leads David Banisar, a policy analyst at Computer Professionals for Social Responsibility, to worry about what he calls "function creep"—the temptation to add features once the national health network is in place. "It could start off as a benign ATM card," Banisar says. "If you add extra functions, it could be a real problem."

Functions may already have begun to creep. Entrepreneurs and health care providers throughout Europe and even in the U.S. are smartening cards. A California firm, LaserCard Systems, recently demonstrated to Congress a four-megabyte card with an optical memory that lets 2,000 pages of encrypted text or multiple images be written on it. In California the Loma Linda School of Medicine has used the card to store a photograph and an x-ray of a patient and the voice of a physician dictating diagnostic information.

Ultimately, social and economic pressure, the American Civil Liberties Union contends, may prove more dangerous than any lapse in a new technology. The Clinton health bill sets out general principles about patient rights to inspect and guard medical data. But most people have little choice about whether to let someone else see a record if such perusal is a condition of receiving benefits or a job. To remedy the problem, Goldman would like to see the establishment of "fire walls"—narrow definitions of the information that any corporation or institution can ask an individual to release.

The brave new world of computerized medical data has already inspired a number of legislative initiatives. Congress may try to bring order to the inconsistent mix of state laws that now govern access to medical records. One problem is that most of the state legislation enacted to date has failed to deal with access to health records by police, the courts, the media and other so-called secondary users. The Federal Privacy Act, a 1974 law, pertains only to medical information gathered by federal agencies—a Veterans Administration hospital, for example. Today access to medical records varies from state to state. "Records in one state may offer a reasonable level of protection, but if you carry them across the river you lose the protection," says Robert M. Gellman, an aide to Representative Gary Condit of California, who held hearings in November 1993 on a bill being written to cover health care privacy.

The Clinton health proposal calls for a 15-member privacy panel to advise the National Health Board. Advocacy groups want to broaden the debate to include more than just medical records and have welcomed a Privacy Protection Commission proposed by Senator Paul Simon of Illinois. Other legislators are also thinking about what a national health plan means for privacy. Besides the bill from Condit, Senator Christopher Bond of Missouri proposed a law in September 1993 that would regulate electronic health care records.

Privacy for medical records may finally have achieved the national standing it deserves. A poll conducted in 1993 by Louis Harris & Associates for Equifax, a credit-reporting agency, revealed that 85 percent of Americans attached great importance to protecting medical records under health care reform. Confidence in the Clinton plan may hinge on assuring people about the integrity of what is told to a physician in confidence. *—Gary Stix*

President Bill Clinton displays the health card.

Civil Liberties in Cyberspace

When does hacking turn from an exercise of civil liberties into crime?

by Mitchell Kapor

On March 1, 1990, the U.S. Secret Service raided the offices of Steve Jackson, an entrepreneurial publisher in Austin, Tex. Carrying a search warrant, the authorities confiscated computer hardware and software, the drafts of his about-to-be-released book and many business records of his company, Steve Jackson Games. They also seized the electronic bulletin-board system used by the publisher to communicate with customers and writers, thereby seizing all the private electronic mail on the system.

The Secret Service held some of the equipment and material for months, refusing to discuss their reasons for the raid. The publisher was forced to reconstruct his book from old manuscripts, to delay filling orders for it and to lay off half his staff. When the warrant application was finally unsealed months later, it confirmed that the publisher was never suspected of any crime.

Steve Jackson's legal difficulties are symptomatic of a widespread problem. During the past several years, dozens of individuals have been the subject of similar searches and seizures. In any other context, this warrant might never have been issued. By many interpretations, it disregarded the First and Fourth Amendments to the U.S. Constitution, as well as several existing privacy laws. But the government proceeded as if civil liberties did not apply. In this case, the government was investigating a new kind of crime—computer crime.

The circumstances vary, but a disproportionate number of cases share a common thread: the serious misunderstanding of computer-based communication and its implications for civil liberties. We now face the task of adapting our legal institutions and societal expectations to the cultural phenomena that even now are springing up from communications technology.

MITCHELL KAPOR founded Lotus Development Corporation in 1982 and is the designer of Lotus 1-2-3, a spreadsheet computer program. He is now the founder and chairman of ON Technology, Inc., which develops software for the Apple Macintosh computer. Kapor holds a B.A. in cybernetics from Yale University and an M.A. in psychology from Beacon College. In 1986 and 1987 he was a visiting scientist at the Massachusetts Institute of Technology's Center for Cognitive Science. He is co-founder and president of the Electronic Frontier Foundation, a nonprofit organization concerned with the application of civil liberties to new electronic media, and is a member of the Computer Science and Telecommunications Board of the National Research Council.

Our society has made a commitment to openness and to free communication. But if our legal and social institutions fail to adapt to new technology, basic access to the global electronic media could be seen as a privilege, granted to those who play by the strictest rules, rather than as a right held by anyone who needs to communicate. To ensure that these freedoms are not compromised, a group of computer experts, including me, founded the Electronic Frontier Foundation (EFF) in 1990.

In many respects, it was odd that Steve Jackson Games got caught up in a computer crime investigation at all. The company publishes a popular, award-winning series of fantasy role-playing games, produced in the form of elaborate rule books. The raid took place only because law enforcement officials misunderstood the technologies—computer bulletin-board systems (BBSs) and on-line forums—and misread the cultural phenomena that those technologies engender.

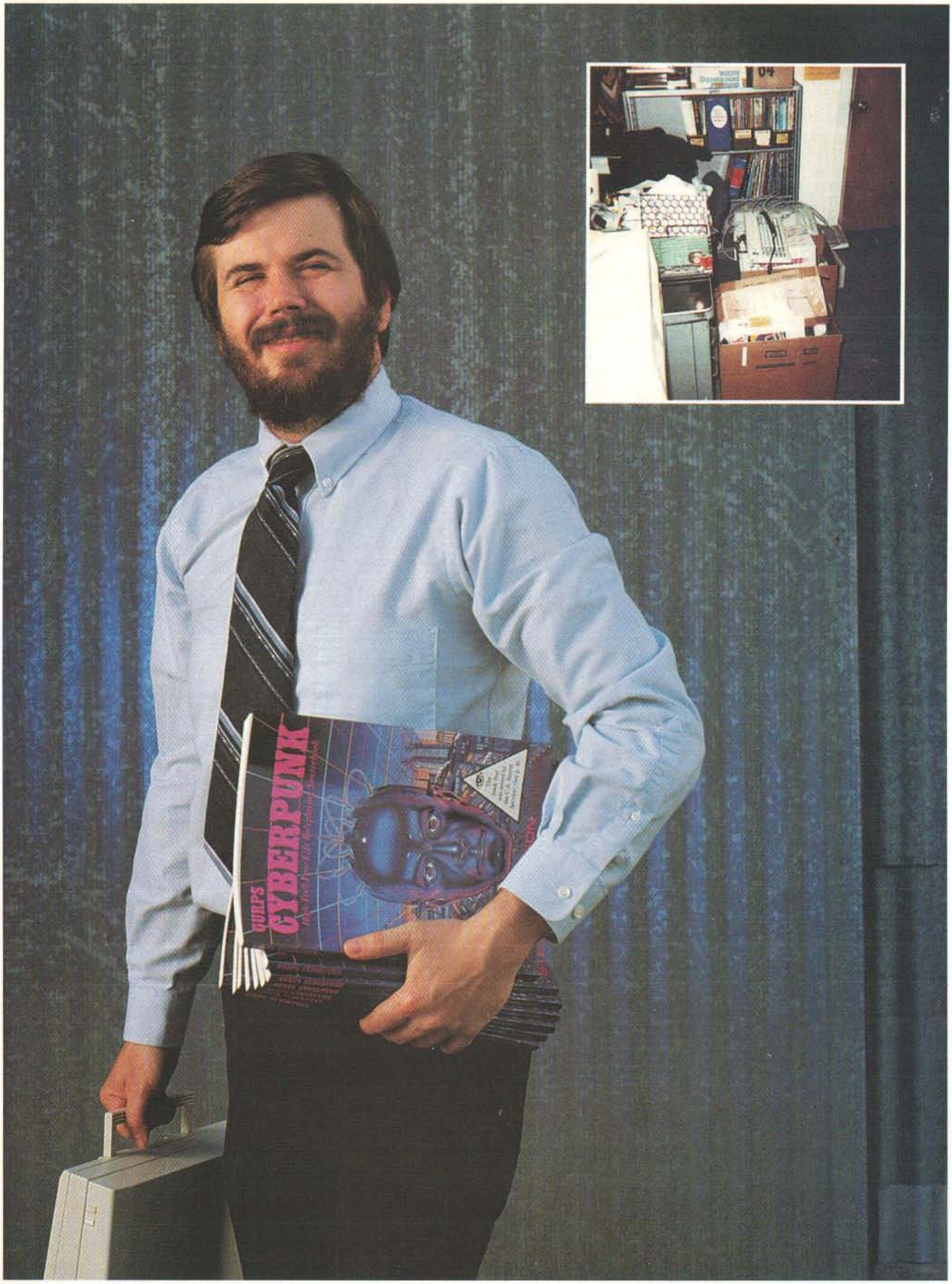
Like a growing number of businesses, Steve Jackson Games operated an electronic bulletin board to facilitate contact between players of its games and their authors. Users of this bulletin-board system dialed in via modem from their personal computers to swap strategy tips, learn about game upgrades, exchange electronic mail and discuss games and other topics.

Law enforcement officers apparently became suspicious when a Steve Jackson Games employee—on his own time and on a BBS he ran from his house—made an innocuous comment about a public domain protocol for transferring computer files called Kermit. In addition, officials claimed that at one time the employee had had on an electronic bulletin board a copy of *Phrack*, a widely disseminated electronic publication, that included information they believed to have been stolen from a BellSouth computer.

The law enforcement officials interpreted these facts as unusual enough to justify not only a search and seizure at the employee's residence but also the search of Steve Jackson Games and the seizure of enough equipment to disrupt the business seriously. Among the items confiscated were all the hard copies and electronically stored copies of the manuscript of a rule book for a role-playing game called GURPS Cyberpunk, in which inhabitants of so-called cyberspace invade corporate and government computer systems and steal sensitive data. Law enforcement agents regarded the book, in the words of one, as "a handbook for computer crime."

A basic knowledge of the kinds of computer intrusion that are technically possible would have enabled the agents to see that GURPS Cyberpunk was nothing more than a science fiction creation and that Kermit was simply a legal, frequently used computer program. Unfortunately, the agents assigned to investigate computer crime did not know what—if anything—was evidence of criminal activity. Therefore, they intruded on a small business without a reasonable basis for believing that a crime had been committed and conducted a search and seizure without looking for particular evidence, in violation of the Fourth Amendment of the Constitution.

Reprinted from the September 1991 issue



GAME PUBLISHER Steve Jackson maintains that his civil rights were violated when the Secret Service seized computers, business records and manuscripts. As a "receipt," Jackson says agents left a Polaroid of his boxed possessions (*inset*).

Searches and seizures of such computer systems affect the rights of not only their owners and operators but also the users of those systems. Although most BBS users have never been in the same room with the actual computer that carries their postings, they legitimately expect their electronic mail to be private and their lawful associations to be protected.

The community of bulletin-board users and computer networkers may be small, but precedents must be understood in a greater context. As forums for debate and information exchange, computer-based bulletin boards and conferencing systems support some of the most vigorous exercise of the First Amendment freedoms of expression and association that this country has ever seen. Moreover, they are evolving rapidly into large-scale public information and communications utilities.

These utilities will probably converge into a digital national public network that will connect nearly all homes and businesses in the U.S. This network will serve as a main conduit for commerce, learning, education and entertainment in our society, distributing images and video signals as well as text and voice. Much of the content of this network will be private messages serving as "virtual" town halls, village greens and coffeehouses, where people post their ideas in public or semipublic forums.

Yet there is a common perception that a defense of electronic civil liberties is somehow opposed to legitimate concerns about the prevention of computer crime. The conflict arises, in part, because the popular hysteria about the technically sophisticated youths known as hackers has drowned out reasonable discussion.

Perhaps inspired by the popular movie *WarGames,* the general public began in the 1980s to perceive computer hackers as threats to the safety of this country's vital computer systems. But the image of hackers as malevolent is purchased at the price of ignoring the underlying reality—the typical teenage hacker is simply tempted by the prospect of exploring forbidden territory. Some are among our best and brightest technological talents: hackers of the 1960s and 1970s, for example, were so driven by their desire to master, understand and produce new hardware and software that they went on to start companies called Apple, Microsoft and Lotus.

How do we resolve this conflict? One solution is ensure that our scheme of civil and criminal laws provides sanctions in proportion to the offenses. A system in which an exploratory hacker receives more time in jail than a defendant convicted of assault violates our sense of justice. Our legal tradition historically has shown itself capable of making subtle and not so subtle distinctions among criminal offenses.

There are, of course, real threats to network and system security. The qualities that make the ideal network valuable—its popularity, its uniform commands, its ability to handle financial transactions and its international access—also make it vulnerable to a variety of abuses and accidents. It is certainly proper to hold hackers accountable for their offenses, but that accountability should never entail denying defendants the safeguards of the Bill of Rights, including the rights to free expression and association and to freedom from unreasonable searches and seizures.

If our legal and social institutions fail to adapt, access to electronic media could be a privilege, rather than a right.

We need statutory schemes that address the acts of true computer criminals (such as those who have created the growing problem of toll and credit-card fraud) while distinguishing between those criminals and hackers whose acts are most analogous to noncriminal trespass. And we need educated law enforcement officials who will be able to recognize and focus their efforts on the real threats.

The question then arises: How do we help our institutions, and perceptions, adapt? The first step is to articulate the kinds of values we want to see protected in the electronic society we are now shaping and to make an agenda for preserving the civil liberties that are central to that society. Then we can draw on the appropriate legal traditions that guide other media. The late Ithiel de Sola Pool argued in his influential book *Technologies of Freedom* that the medium of digital communications is heir to several traditions of control: the press, the common carrier and the broadcast media.

The freedom of the press to print and distribute is explicitly guaranteed by the First Amendment. This freedom is somewhat limited, particularly by laws governing obscenity and defamation, but the thrust of First Amendment law, especially in this century, prevents the government from imposing "prior restraint" on publications.

Like the railroad networks, the telephone networks follow common-carrier principles—they do not impose content restrictions on the "cargo" they carry. It would be unthinkable for the telephone company to monitor our calls routinely or cut off conversations because the subject matter was deemed offensive.

Meanwhile the highly regulated broadcast media are grounded in the idea, arguably mistaken, that spectrum scarcity and the pervasiveness of the broadcast media warrant government allocation and control of access to broadcast frequencies (and some control of content). Access to this technology is open to any consumer who can purchase a radio or television set, but it is nowhere near as open for information producers.

Networks as they now operate contain elements of publishers, broadcasters, bookstores and telephones, but no one model fits. This hybrid demands new thinking or at least a new application of the old legal principles. As hybrids, computer networks also have some features that are unique among the communications media. For example, most conversations on bulletin boards, chat lines and conferencing systems are both public and private at once. The electronic communicator speaks to a group of individuals, only some of whom are known personally, in a discussion that may last for days or months.

But the dissemination is controlled, because the membership is limited to the handful of people who are in the virtual room, paying attention. Yet the result may also be "published"—an archival textual or voice record can be automatically preserved, and newcomers can read the backlog. Some people tend to equate on-line discussions with party (or party-line) conversations, whereas others compare them to newspapers and still others think of citizens band radio.

In this ambiguous context, free-speech controversies are likely to erupt. Last year an outcry went up against the popular Prodigy computer service, a joint venture of IBM and Sears, Roebuck and Co. The problem arose because Prodigy management regarded their service as essentially a "newspaper" or "magazine," for which a hierarchy of editorial control is appropriate. Some of Prodigy's customers, in contrast, regarded the service as more of a forum or meeting place.

When users of the system tried to

protest Prodigy's policy, its editors responded by removing the discussion. When the protestors tried to use electronic mail as a substitute for electronic assembly, communicating through large mailing lists, Prodigy placed a limit on the number of messages each individual could send.

The Prodigy controversy illustrates an important principle that belongs on any civil liberties agenda for the future: freedom-of-speech issues will not disappear simply because a service provider has tried to impose a metaphor on its service. Subscribers sense, I believe, that freedom of speech on the networks is central for individuals to use electronic communications. Science fiction writer William Gibson once remarked that "the street finds its own uses for things." Network service providers will continue to discover that their customers will always find their own best uses for new media.

Freedom of speech on networks will be promoted by limiting content-based regulations and by promoting competition among providers of network services. The first is necessary because governments will be tempted to restrict the content of any information service they subsidize or regulate. The second is necessary because market competition is the most efficient means of ensuring that needs of network users will be met.

The underlying network should essentially be a "carrier"—it should operate under a content-neutral regime in which access is available to any entity that can pay for it. The information and forum services would be "nodes" on this network. (Prodigy, like GEnie and CompuServe, currently maintains its own proprietary infrastructure, but a future version of Prodigy might share the same network with services like CompuServe.)

Each service would have its own unique character and charge its own rates. If a Prodigy-like entity correctly perceives a need for an electronic "newspaper" with strong editorial control, it will draw an audience. Other, less hierarchical services will share the network with that "newspaper" yet find their own market niches, varying by format and content.

The prerequisite for this kind of competition is a carrier capable of high-bandwidth traffic that is accessible to individuals in every community. Like common carriers, these network carriers should be seen as conduits for the distribution of electronic transmissions. They should not be allowed to change the content of a message or to discriminate among messages.

This kind of restriction will require shielding the carriers from legal liabilities for libel, obscenity and plagiarism. Today the ambiguous state of liability law has tempted some computer network carriers to reduce their risk by imposing content restrictions. This could be avoided by appropriate legislation. Our agenda requires both that the law shield carriers from liability based on content and that carriers not be allowed to discriminate.

Network service providers will continue to discover that their customers will always find their own best uses for the new media.

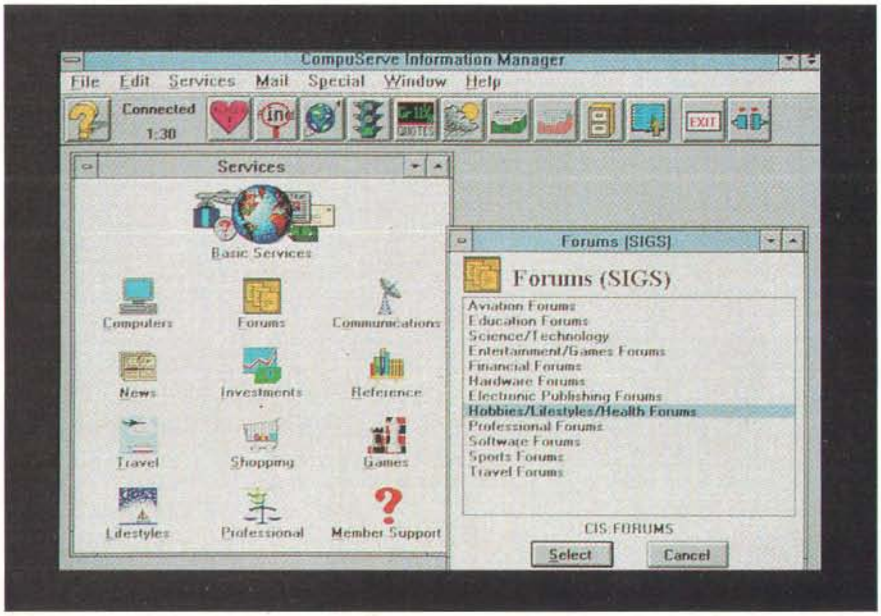
All electronic "publishers" should be allowed equal access to networks. Ultimately, there could be hundreds of thousands of these information providers, as there are hundreds of thousands of print publishers today. As "nodes," they will be considered the conveners of the environments within which on-line assembly takes place.

None of the old definitions will suffice for this role. For example, to safeguard the potential of free and open inquiry, it is desirable to preserve each electronic publisher's control over the general flow and direction of material under his or her imprimatur—in effect, to give the "sysop," or system operator, the prerogatives and protections of a publisher.

But it is unreasonable to expect the sysop of a node to review every message or to hold the sysop to a publisher's standard of libel. Message traffic on many individually owned services is already too great for the sysop to review. We can only expect the trend to grow. Nor is it appropriate to compare nodes to broadcasters (an analogy likely to lead to licensing and content-based regulation). Unlike the broadcast media, nodes do not dominate the shared resource of a public community, and they are not a pervasive medium. To take part in a controversial discussion, a user must actively seek entry into the appropriate node, usually with a subscription and a password.

Anyone who objects to the content of a node can find hundreds of other systems where they might articulate their ideas more freely. The danger is if choice is somehow restricted: if all computer networks in the country are restrained from allowing discussion on particular subjects or if a publicly sponsored computer network limits discussion.

This is not to say that freedom-of-speech principles ought to protect all electronic communications. Exceptional cases, such as the BBS used primarily to traffic in stolen long-distance access codes or credit-card numbers, will always arise and pose problems of civil and criminal liability. We know that electronic freedom of speech, whether in public or private systems, cannot be



INFORMATION AND FORUM SERVICE operated by CompuServe offers this menu on a user's computer screen. Defining freedom of speech for users of such networks is an unresolved civil liberties issue.

absolute. In face-to-face conversation and printed matter today, it is commonly agreed that freedom of speech does not cover the communications inherent in criminal conspiracy, fraud, libel, incitement to lawless action and copyright infringement.

If there are to be limits on electronic freedom of speech, what precisely should those limits be? One answer to this question is the U.S. Supreme Court's 1969 decision in *Brandenburg* v. *Ohio*. The court ruled that no speech should be subject to prior restraint or criminal prosecution unless it is intended to incite and is likely to cause imminent lawless action.

In general, little speech or publication falls outside of the protections of the Brandenburg case, since most people are able to reflect before acting on a written or spoken suggestion. As in traditional media, any on-line messages should not be the basis of criminal prosecution unless the Brandenburg standard is met.

Other helpful precedents include cases relating to defamation and copyright infringement. Free speech does not mean one can damage a reputation or appropriate a copyrighted work without being called to account for it. And it probably does not mean that one can release a virus across the network in order to "send a message" to network subscribers. Although the distinction is trickier than it may first appear, the release of a destructive program, such as a virus, may be better analyzed as an act rather than as speech.

Following freedom of speech on our action agenda is freedom from unreasonable searches and seizures. The Steve Jackson case was one of many cases in which computer equipment and disks were seized and held—sometimes for months—often without a specific charge being filed. Even when only a few files were relevant to an investigation, entire computer systems, including printers, have been removed with their hundreds of files intact.

Such nonspecific seizures and searches of computer data allow "rummaging," in which officials browse through private files in search of incriminating evidence. In addition to violating the Fourth Amendment requirement that searches and seizures be particular, these searches often run afoul of the Electronic Communications Privacy Act of 1986. This act prohibits the government from seizing or intercepting electronic communications without proper authorization. They also contravene the Privacy Protection Act of 1980, which prohibits the government from searching the offices of publishers for documents, including materials that are electronically stored.

We can expect that law enforcement agencies and civil libertarians will agree over time about the need to establish procedures for searches and seizures of particular computer data and hardware. Law enforcement officials will have to adhere to guidelines in the above statutes to achieve Fourth Amendment "particularity" while maximizing the efficiency of their searches. They also will have to be trained to make use of software tools that allow searches for particular files or particular information within files on even the most capacious hard disk or optical storage device.

Implementing a civil liberties agenda for computer networks will require participation by technically trained people.

Still another part of the solution will be law enforcement's abandonment of the myth of the clever criminal hobbyist. Once law enforcement no longer assumes worst-case behavior but looks instead for real evidence of criminal activity, its agents will learn to search for and seize only what they need.

Developing and implementing a civil liberties agenda for computer networks will require increasing participation by technically trained people. Fortunately, there are signs that this is beginning to happen. The Computers, Freedom and Privacy Conference, held in San Francisco, together with electronic conferences on the WELL (Whole Earth 'Lectronic Link) and other computer networks, has brought law enforcement officials, supposed hackers and interested members of the computer community together in a spirit of free and frank discussion. Such gatherings are beginning to work out the civil liberties guidelines for a networked society.

There is general agreement, for example, that a policy on electronic crime should offer protection for security and privacy on both individual and institutional systems. Defining a measure of damages and setting proportional punishment will require further good-faith deliberations by the community involved with electronic freedoms, including the Federal Bureau of Investigation, the Secret Service, the bar associations, technology groups, telephone companies and civil libertarians. It will be especially important to represent the damage caused by electronic crime accurately and to leave room for the valuable side of the hacker spirit: the interest in increasing legitimate understanding through exploration.

We hope to see a similar emerging consensus on security issues. Network systems should be designed not only to provide technical solutions to security problems but also to allow system operators to use them without infringing unduly on the rights of users. A security system that depends on wholesale monitoring of traffic, for example, would create more problems than it would solve.

Those parts of a system where damage would do the greatest harm—financial records, electronic mail, military data—should be protected. This involves installing more effective computer security measures, but it also means redefining the legal interpretations of copyright, intellectual property, computer crime and privacy so that system users are protected against individual criminals and abuses by large institutions. These policies should balance the need for civil liberties against the need for a secure, orderly, protected electronic society.

As we pursue that balance, of course, confrontations will continue to take place. In May 1991, Steve Jackson Games, with the support of the EFF, filed suit against the Secret Service, two individual Secret Service agents, an assistant U.S. attorney and others.

The EFF is not seeking confrontation for its own sake. One of the realities of our legal system is that one often has to fight for a legal or constitutional right in the courts in order to get it recognized outside the courts. One goal of the lawsuit is to establish clear grounds under which search and seizure of electronic media is "unreasonable" and unjust. Another is to establish the clear applicability of First Amendment principles to the new medium.

But the EFF's agenda extends far beyond litigation. Our larger agenda includes sponsoring a range of educational initiatives aimed at the public's general lack of familiarity with the technology and its potential. That is why there is an urgent need for technologically knowledgeable people to take part in the public debate over communications policy and to help spread their understanding of these issues. Fortunately, the very technology at stake—electronic conferencing—makes it easier than ever before to get involved in the debate.

Tap Dance

Keeping communications networks safe for bugging

The Federal Bureau of Investigation was once fond of pointing out that "it always gets its man." One of its most heavily used techniques—as in the murder trial of mob boss John Gotti—is the trusty wiretap. But the bureau fears that new digital communications systems could leave its agents listening to an uninformative cacophony. In the spring of 1992 the U.S. Department of Justice proposed legislation requiring makers of electronic communications equipment to ensure that their equipment could be tapped.

While civil libertarians rose up in outrage, many communications engineers scratched their heads in puzzlement. The idea that digital technology could make it impossible to intercept communications appeared to be somewhere between incomprehensible and flat wrong. Soon after, participants in the Telecom Digest, an international on-line discussion of communications-related issues, outlined how existing digital-switching technology could be used to perform completely undetectable taps.

One technique is simply to turn every call to or from a tapped phone into a conference call with law enforcement on the third line. Another way would be to allow officers access to "maintenance ports" normally used to monitor phone line quality. "Maybe the FBI doesn't know how to do it yet," asserts Michael A. Schwartz, an engineer at Bellcore, but "hundreds of thousands of people" outside the bureau know how to tap digital phone lines. It's a question of getting the right equipment, he says. Special electronics capable of tapping facsimile transmissions have been commercially available for years, for example.

The bureau is not pleased at having its technical prowess impugned. William A. Bayse, the FBI's assistant director for technical services, snaps, "It's not worth having this conversation" at a reporter who asks him about the claims made by communications engineers. Bayse contends that few facilities are available for tapping digital conversations, and, furthermore, he declares, digital phone standards are still in flux—not only must would-be listeners get their taps in place, they must also figure out in just what form their information is being transmitted.

Bayse has a point. Tapping digits is not as easy as hooking up the time-honored "alligator clips"—those tiny spring-loaded connectors that clamp onto a piece of copper wire and let anyone with earphones or a tape recorder listen in. These days it is as likely to be the singularly uninteresting sound of bits as an analog human voice. Not only must a digital wiretap convert bits to sound, it must also cope with the possibility that a digital line is carrying anywhere from two to dozens of conversations and strip out only the information authorized by warrant. At some point when optical fibers replace copper telephone wires, tappers could be forced to invest in the extraordinarily finicky gadgetry required to siphon off photons without being detected.

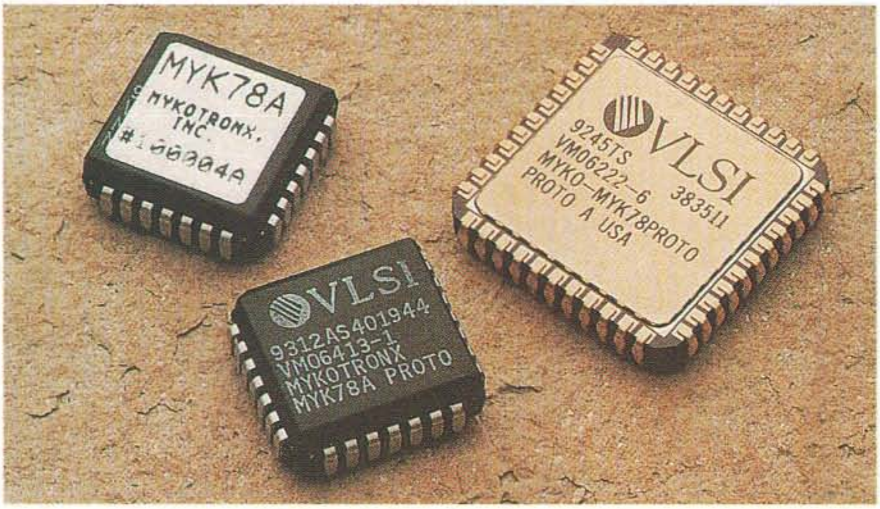
Faced with such an uncertain future, law enforcement officials might well prefer to get their taps directly from the telephone switching system, but engineers will guard their system jealously. Schwartz contends that adding convenient wiretapping to the long list of other features for call routing and tracing on a digital switch at best would be expensive and at worst could make the phone system unreliable.

Cost is also a sensitive issue. Under the proposed legislation, telephone companies and equipment makers would be allowed to charge customers for the privilege of being tappable. FBI director William Sessions has estimated the cost at a mere $250 million to $300 million.

And the U.S. telephone system is only the beginning of the problem. The FBI's proposal applied equally to private branch exchanges (internal corporate telephone systems), local-area networks, computer bulletin boards and even the Internet, which links computers across the U.S. and throughout the world. All these systems, according to the FBI's bill, would have to conform to (potentially unpublished) ease-of-tapping standards promulgated by the Federal Communications Commission. Violators would face fines up to $10,000 a day.

Computer manufacturers have been working hard to make their hardware and software more secure and to build in encryption and other techniques that will safeguard information from prying eyes, says John Podesta, a former congressional staffer who helped to draft the Electronic Communications Privacy Act of 1986. Consequently, computer makers are not welcoming the idea of increased "porousness."

After first discussing the issue only with telephone companies, the FBI met privately with some computer companies. It later revised the bill—now named the Digital Telephony and Communications Privacy Improvement Act of 1994—to limit surveillance to public networks and leave out private branch exchanges and private corporate computer networks. And in discussions with the House Judiciary Committee's subcommittee on civil rights, the FBI and opponents of the legislation are trying to reach a compromise. *—Paul Wallich*



CLIPPER CHIP is the FBI's choice for making coded information on a computer network tappable. But Matthew Blaze of AT&T Bell Laboratories recently reported that the chip will not work as expected because of technical flaws.

The Risks of Software

Programming bugs have disrupted telephone service and delayed shuttle launches. An inherent uncertainty in reliability may mean limiting a computer's role, especially in systems where software is critical for safety

by Bev Littlewood and Lorenzo Strigini

Most of us have experienced some kind of problem related to computer failure: a bill mailed in error or a day's work destroyed by some mysterious glitch in a desktop computer. Such nuisances, often caused by software faults, or "bugs," are merely inconvenient when compared with the consequences of computer failures in critical systems. Software bugs have caused a series of large-scale outages of telephone service in the U.S. A software problem may have prevented the Patriot missile system from tracking the Iraqi Scud missile that killed 28 American soldiers during the Gulf War. Indeed, software faults are generally more insidious and much more difficult to handle than are physical defects.

The problems essentially arise from complexity, which increases the possibility that design faults will persist and emerge in the final product. Conventional engineering has made great strides in the understanding and control of physical problems. Although design faults are sometimes present in material products that do not contain computers, the relative simplicity of such machines has made design reliability less serious than it has become for software. The concerns expressed here, incidentally, go far beyond exotic military and aerospace products. Complex software is finding critical roles in more mundane areas, such as four-wheel steering and antilock braking in automobiles.

BEV LITTLEWOOD and LORENZO STRIGINI collaborate as members of the project Predictably Dependable Computing Systems (PDCS), a research effort on computer dependability that brings together researchers from European countries. Littlewood received his Ph.D. in computer science and statistics from the City University in London, where he is the director of the Centre for Software Reliability. Strigini, a researcher at the Institute for Information Processing (IEI) of Italy's National Research Council (CNR), received his *laurea* in electronic engineering from the University of Pisa. The work of the authors in this field was partially funded by the Commission of the European Communities as part of the PDCS project. The authors thank the other workers in PDCS, discussions with whom have been valuable in forming the views expressed here.

In this article we examine some major reasons for the uncertainty concerning software reliability and argue that our ability to measure it falls far short of the levels that are sometimes required. In critical systems, such as the safety systems of a dangerous chemical plant, it may be that the appropriate level of safety will be guaranteed only if the role of software is limited.

In theory at least, software can be made that is free of defects. Unlike materials and machinery, software does not wear out. All design defects are present from the time the software is loaded into the computer. In principle, these faults could be removed once and for all. Furthermore, mathematical proof should enable programmers to guarantee correctness.

Yet the goal of perfect software remains elusive. Despite rigorous and systematic testing, most large programs contain some residual bugs when delivered. The reason for this is the complexity of the source code. A program of only a few hundred lines may contain tens of decisions, allowing for thousands of alternative paths of execution (programs for fairly critical applications vary between tens and millions of lines of code). A program can make the wrong decision because the particular inputs that triggered the problem had not been used during the test phase, when defects could be corrected. The situation responsible for such inputs may even have been misunderstood or unanticipated: the designer either "correctly" programmed the wrong reaction or failed to take the situation into account altogether. This type of bug is the most difficult to eradicate.

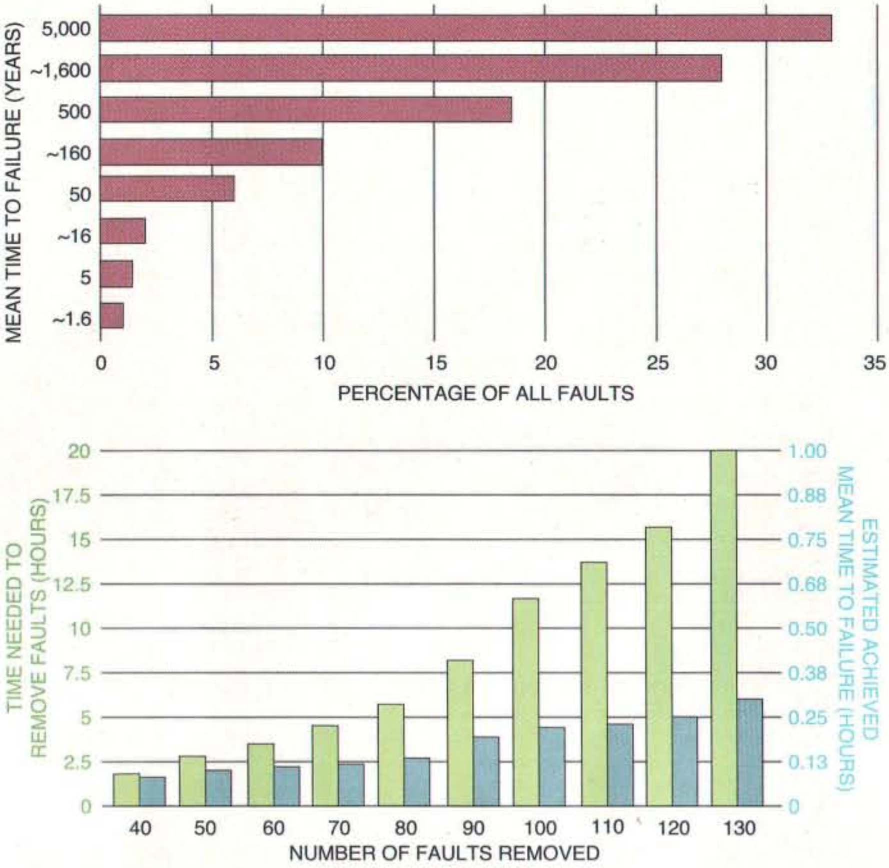
PATRIOT MISSILE streaks over Tel Aviv to intercept an incoming Iraqi Scud mis-

In addition, specifications often change during system development, as the intended purpose of the system is modified or becomes better understood. Such changes may have implications that ripple through all parts of a system, making the previous design inadequate. Furthermore, real use may still differ from intended purpose. Failures of Patriot missiles to intercept Scud missiles have been attributed to an accumulation of inaccuracies in the internal time-keeping of a computer. Yet the computer was performing according to specifications: the system was meant to be turned off and restarted often enough for the accumulated error never to become dangerous. Because the system was used in an unintended way, a minor inaccuracy became a serious problem.

The intrinsic behavior of digital systems also hinders the creation of completely reliable software. Many physical systems are fundamentally continuous in that they are described by "well-behaved" functions—that is, very small changes in stimuli produce very small differences in responses. In contrast, the smallest possible perturbation to the state of a digital computer (changing a bit from 0 to 1, for instance) may produce a radical response. A single incorrect character in the specification of a control program for an Atlas rocket, carrying the first U.S. interplanetary spacecraft, *Mariner 1,* ultimately caused the vehicle to veer off course. Both rocket and spacecraft had to be destroyed shortly after launch.

In all other branches of engineering, simplicity and gradual change constitute the main elements of trustworthy design. But in software engineering the unprecedented degrees of novelty and

sile during the 1991 Gulf War. On some occasions the software controlling the Patriot's tracking system may have failed, preventing the missiles from locating and destroying their Scud targets. One such failure led to the deaths of 28 U.S. soldiers.



SOFTWARE FAULTS persist even in well-debugged programs. Edward N. Adams of IBM found that bugs that remained in a system were primarily "5,000-year" bugs—that is, each of them would produce a failure only once in 5,000 years (*top*). Such faults make debugging an exercise in diminishing returns: in the test of a military command-and-control system (*bottom*), the time needed to remove the bugs begins to outpace by far the resulting improvement in the estimated reliability, measured in terms of estimated achieved mean time to failure. For visual clarity, the graphs have been plotted on different time scales.

flexibility that programming affords tempt workers to ignore these principles. Entirely new applications can be designed with apparent ease, giving a false sense of security to developers and clients who are not familiar with problems specific to software. Even the addition of novel features to a program may produce unexpected changes in existing features.

The problems of embedding complex decision rules in a design and forecasting the behavior of complex discontinuous systems are not limited to software. Designers of highly complex digital integrated circuits encounter similar problems. Software, however, is still the predominant medium for embodying extremely complex, specialized decision rules.

In addition to unintentional design bugs, flaws deliberately introduced to compromise a system can cause unacceptable system behavior. The issue of computer security, privacy and encryption requires special consideration that lies beyond the scope of this article [see "Achieving Electronic Privacy," by David Chaum; SCIENTIFIC AMERICAN, August 1992].

Given that perfect software is a practical impossibility, how can we decide whether a program is as reliable as it is supposed to be? First, safety requirements must be chosen carefully to reflect the nature of the application. These requirements can vary dramatically from one application to another. For example, the U.S. requires that its new air-traffic control system cannot be unavailable for more than three seconds a year. In civilian airliners, the probability of certain catastrophic failures must be no worse than 10^{-9} per hour.

In setting reliability requirements for computers, we must also take into account any extra benefits that a computer may produce, because not using a particular system may itself incur harm. For example, military aircraft are by necessity much more dangerous to fly than are civilian ones. Survival in combat depends on high performance, which forbids conservative design, and a new computer system may improve the airplane's chances even if it is less safe than computers used in commercial airplanes. Similarly, in the design of a fly-by-wire civilian aircraft, such as the Airbus A320 or the Boeing 777, the possibility that software may cause accidents has to be weighed against the likelihood that it may avoid some mishaps that would otherwise be caused by pilot error or equipment failure.

We believe that there are severe restrictions on the levels of confidence that one can justifiably place in the reliability of software. To explain this point of view, we need to consider the different sources of evidence that support confidence in software. The most obvious is testing: running the program, directly observing its behavior and removing bugs whenever they show up. In this process the reliability of the software will grow, and the data collected can now generally be used, via sophisticated statistical extrapolation techniques, to obtain accurate measures of how reliable the program has become.

Unfortunately, this approach works only when the reliability requirements are fairly modest (say, in the range of one failure every few years) when compared with the requirements often set for critical applications. To have confidence at a level such as 10^{-9} failure per hour, we would need to execute the program for very many multiples of 10^9 hours, or 100,000 years. Clearly, this task is not possible. In the time spans for which it is feasible to test, assurance of the safety would fall many orders of magnitude short of what is needed.

The problem here is a law of diminishing returns. When we continue debugging a program for a very long time, eventually the bugs found are so "small" that fixing them has virtually no effect on the overall reliability or safety. Edward N. Adams of the IBM Thomas J. Watson Research Center empirically analyzed "bug sizes" over a worldwide data base that involved the equivalent use of thousands of years of a particular software system.

The most extraordinary discovery was that about a third of all bugs found were "5,000-year" bugs: each of them produced a failure only about once in 5,000 years of execution (the rates from other bugs varied by several orders of magnitude). These rare bugs made up a sizable portion of all faults because bugs that caused higher failure rates were encountered, and so removed, earlier. Eventually, only the 5,000-year bugs will make the system unreliable, and removal of one of these will bring negligible improvement in reliability.

Extrapolating from testing and debugging also implies an unsubstantiated assumption—namely, that a bug, once encountered, is simply corrected. In reality, an attempt to fix a bug sometimes fails. It may even introduce an entirely novel fault. Because nothing would be known about the new bug, its effect on the reliability of the system would be unbounded. In particular, the system might not even be as reliable as it was before the bug was found.

Therefore, a prudent course would be to discount completely the history prior to the last failure. This precaution, critically important in situations that involve safety, would require an evaluator to treat the software after the last fix as if it were a completely new program. Only the most recent period of error-free working would influence judgment about the safety of the program. But even this conservative course of action cannot provide much confidence. Our research has shown that under quite plausible mathematical assumptions, there is only about a 50–50 chance that the program will function without failure for the same length of time as it had before.

The problem of estimating safety is actually even more serious. To have any confidence in the numerical results, we must subject the program to situations it might encounter in reality. This approach ensures that inputs causing failures are encountered with the same frequency with which they would in fact arise. In addition, the tester should always be able to decide whether the program's output is actually correct. The problems here are similar to those of designing and implementing the software itself. To construct an accurate test environment, we need to be sure that we have thought of all circumstances that the software will meet. Just as the unexpected often defeats us in system design, so it is in test design. It would be wise to retain an element of skepticism about the representativeness of the testing and thus about the accuracy of the figures.

The problem in demonstrating extreme reliability or safety for any individual piece of software is simply lack of the necessary knowledge. For complex software, the unpalatable truth seems to be that there are severe limitations to the confidence one can place in a program. Merely observing a program's behavior is not the way to be sure that it will function properly for 100,000 years. How else might we acquire such confidence?

An obvious prerequisite for high reliability is that software be built with methods that are likely to achieve reliability. One method uses "formal" techniques, which rely on mathematical proofs to guarantee that a program will function according to specification. Indeed, formal techniques have become a topic of wide interest. Such methods, though currently limited by practical problems in their scope of application, can effectively avoid programming errors arising in the translation from the specification to the actual program.

Unfortunately, specifications must also be formal statements. In other words, the user's needs would have to be expressed in a mathematical language. That task is not simple: it requires a careful choice of those aspects of the real world to be described in the formal language and an understanding of both the detailed practical problems of the application and of the formal language. Errors would likely be introduced during this process, and we could not reasonably claim the program would never fail.

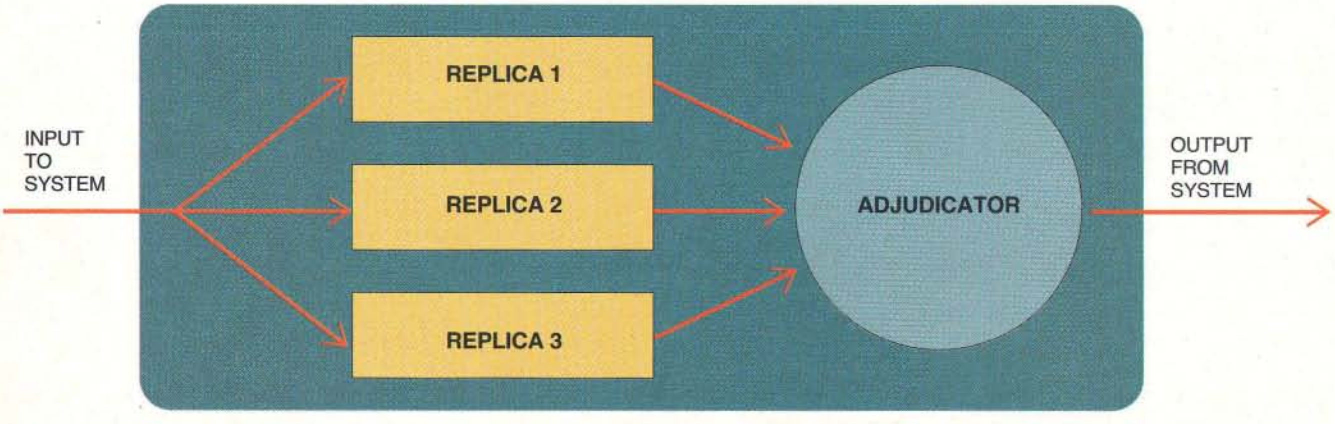
Another method now widely used (in avionic and railroad control applications, for instance) to achieve high reliability is fault tolerance, or protective redundancy. A typical way of applying redundancy is to have different design teams develop several versions of the program. The hope is that if the teams make mistakes, the errors will be different. Each version of the program provides its "opinion" of the correct output. The outputs pass to an adjudication phase, which produces a single output that would be correct if the majority of versions gave the correct result.

Some evidence exists that such design diversity delivers high reliability in a cost-effective manner. Different design teams, however, may make the same mistakes (perhaps because of commonalities in cultural background) or conceptually different mistakes that happen to make the versions fail on the same fault. The adjudicator would therefore produce incorrect output.

To measure the reliability of fault-tolerant software, it is necessary to gauge the statistical correlation between failures of the different versions. Unfortunately, the task turns out to be as hard as trying to measure the reliability by treating the whole system as a single entity—and we have seen the difficulty of doing that.

So if formal proofs do not enable one to claim that a program will never fail and if fault tolerance cannot guarantee reliability, there seems no choice but to evaluate reliability directly, using methods that are acknowledged to be of limited adequacy. How do the regulatory authorities and software users deal with this uncertainty?

There are three approaches. The first

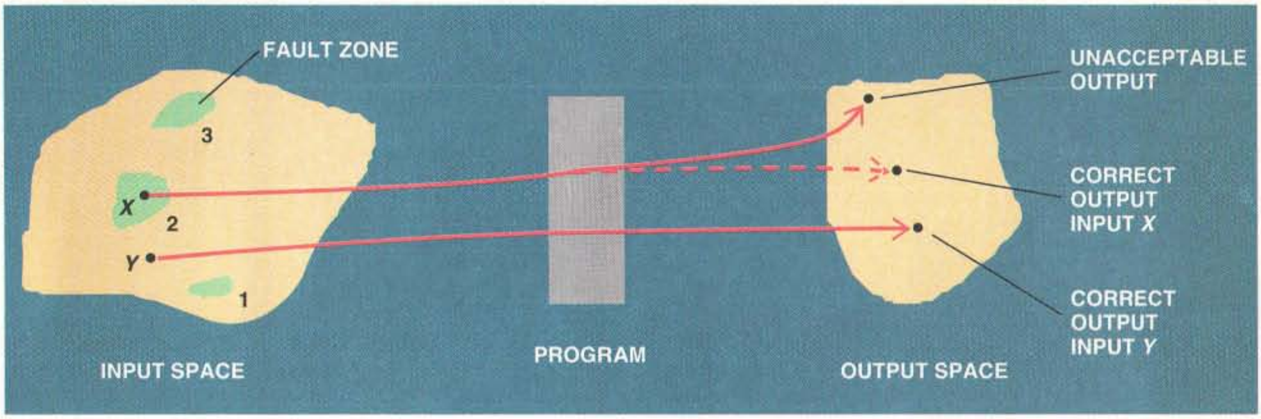


DESIGN DIVERSITY helps to increase the reliability of software systems. Each program version, or replica, is developed independently by different design teams. The "adjudicator" decides the actual output of the system by using, for example, the median value produced by the replicas or the value "voted for" by a majority. The adjudicator could be another redundant system or could consist of noncomputer technology, such as hydraulic actuators.

classifies design-caused failures as "nonquantifiable" errors and avoids specifying requirements for the software. This method is now in fairly wide use. For instance, in civil aviation, the U.S. Federal Aviation Administration Advisory Circular 25.1309-1A describes "acceptable" means for showing compliance with some federal aviation regulations. It states that catastrophic failure conditions (the worst category) must be "so unlikely that they are not anticipated to occur during the entire operational life of all airplanes of one type." The suggested quantitative expression is the probability of failure of not more than 10^{-9} per hour of flight. Software, however, is explicitly excluded from this circular, "because it is not feasible to assess the number or kinds of software errors, if any, that may remain after the completion of system design, development, and test."

The widely used document of the Radio Technical Commission for Aeronautics, RTCA/DO-178A, similarly avoids software measures. The document, which gives guidelines for manufacturers who must seek certification by aviation authorities, explicitly refuses to mandate quantitative terms or methods for evaluating software reliability or safety. Instead the commission regards a correct engineering approach—tight management, thorough reviews and tests, and analysis of previous errors—as more critical than quantitative methods. The basic message of RTCA/DO-178A "is that designers must take a disciplined approach to software: requirements definition, design, development, testing, configuration management and documentation." That is, the best assurance of reliability is to verify that utmost care was used in the design.

How good is such assurance? Arguably, not very good: there is no evidence that superior design and produc-



The Nature of Software Failure

Software occasionally fails because it contains design faults. Some have argued that such failures are systematic—that is, because writing software is a purely logical exercise, there is nothing intrinsically uncertain about it. If enough is known about the inputs, the program's behavior would be completely deterministic. We believe, however, that software failures cannot be mathematically described only in deterministic terms. In fact, we think that describing the nature of software failures requires a probabilistic treatment, just as we use statistics to describe how often, on average, electrical or mechanical devices fail.

To see why, consider all the possible inputs (called the input space) that the software might encounter in its life (*above*). An input for an operation of the software is a set of digital data (numbers) read from the outside world and from information already stored in the computer's memory. In the figure above, the input space is shown in the two dimensions of the printed page, but in practice the space would usually consist of many dimensions.

Here the input space contains three fault zones numbered 1 to 3. Input *x* lies in fault 2, that is, it would cause the program to produce an unacceptable output. On the other hand, the program can successfully execute input *y*, which does not lie in any fault zone.

A program is tested by executing it with many inputs and checking whether the results are correct. If a right answer is produced during testing, it will also be produced whenever the same input is presented. For most programs, testing for all possible inputs would require billions of billions of years—hence, the need to infer failure probabilities from testing on a sample of inputs.

We would like to know when the program will next fail, but that is not possible because of the inherent uncertainty in the process. First, uncertainty arises from the physical mechanisms that determine the succession of inputs (called the trajectory in the input space). We can never be sure which inputs will be selected in the future, and different inputs will have different chances of being selected. Second, we are uncertain about the sizes and locations of the fault regions in the input space. Even if we knew the trajectory, we would still not know when the program would encounter a fault.

Therefore, we must describe our belief about the future failure behavior of the program in terms of probabilities. We might ask what is the probability that we can survive a particular number of inputs before failure. Or we might ask what the probability is that a randomly selected input causes a failure. Both questions can often easily be turned into a time-based measure of reliability—that is, the probability that the program will execute perfectly for a particular length of time.

In conclusion, we are forced to consider the process of successive failures of a program to be just as "random" as that of a hardware device. The use of a probability-based reliability measure is therefore inevitable.

tion methods consistently yield superior products. We cannot even be certain whether the best current methods ever produce sufficient reliability for the more demanding applications.

Eschewing the quantification of software safety poses a serious limitation for many potentially dangerous systems, especially those that require an overall probabilistic risk assessment before operation. Probabilities can be predicted with reasonable accuracy for physical failures caused by stress and wear. But this accuracy cannot be used in efforts to assess the risk that the entire system (that is, hardware and software) will fail if nothing more precise can be said other than that the best effort was made to avoid mistakes. Simply mandating the use of "best practice" does not solve the problem. We hasten to add that it would be foolish to abandon techniques known to improve reliability and safety just because we do not know exactly how much they help. Standards that encourage their use are certainly beneficial, but they do not solve the problem of knowing that the software has the required safety.

The second—and we think, better—approach would require that the system be designed so the role of software in it is not too critical. "Not too critical" here means that the required software reliability is sufficiently modest so the reliability can be demonstrated before the system is deployed. This approach has been taken for the new Sizewell B nuclear reactor in the U.K., where only a 10^{-4} probability of failure on demand is needed from the software-based protection system.

There are well-established methods for limiting the criticality of any one component. For example, an industrial plant whose operations are controlled primarily by computers may be equipped with safety systems that do not depend on any software or other complex design. A safety or backup system usually performs simpler functions than does the main control system, so it can be built more reliably. Safety is possible if the backup systems are completely separated from the main systems. They could be built with different technology or use alternative sensors, actuators and power sources. Then the probability that both primary and backup (or safety) systems will fail simultaneously may be justifiably considered low.

The third approach is simply to accept the current limitations of software and live with a more modest overall system safety. After all, society sometimes demands extremely high safety for what may be irrational reasons. Medical systems are a good example. Surgeons are known to have fairly high failure rates, and it would seem natural to accept a computerized alternative if the device is shown to be as good as or only slightly better than the human physician. Indeed, in the near future robotic surgeons will probably perform operations that are beyond the capabilities of humans.

SIZEWELL B will be the first nuclear reactor in the U.K. to contain both conventional and software-based protection systems for emergency shutdowns. Critics argue that the complexity of the software system—it relies on hundreds of microprocessors and more than 100,000 lines of code—makes it difficult to ensure reactor safety.

The three approaches to regulating software safety may seem rather disappointing. Each sets limits on either the degree of safety in the system or the amount of complexity in the program. Perhaps the only way to learn more about the necessary compromises between safety and complexity is to study the failures (or lack thereof) of software in operation.

Unfortunately, there is a paucity of data from which to fashion statistical predictions. Information on software failure is seldom made public. Companies fear that sharing such knowledge would harm their competitive stance. They worry even more that publishing it would antagonize public opinion. People might see the detection of a software fault as an indication of low production standards, even though it may actually attest to a very thorough procedure applied to very high quality software. But secrecy can only allow expectations of safety to climb to increasingly unrealistic levels. Some investigators have suggested that the government make mandatory the logging and disclosing of failure data in critical software systems. Such regulations would remove the fear that companies volunteering the information would be hurt.

However it is obtained, an extensive collection of data would in time help to quantify the efficacy of different production and validation techniques. The information would help establish more realistic rules for gauging the trustworthiness of software systems. Thus, for software that is not fully tested statistically, the acceptable claims of safety could be tied to explicit upper bounds that would depend on the complexity of the program. Such an approach might allow us to justify claims for the reliability and safety of software beyond what is now believable.

In the meantime, we should remain wary of any dramatic claims of reliability. Considering the levels of complexity that software has made possible, we believe being skeptical is the safest course of action.

FURTHER READING

EVALUATION OF SAFETY-CRITICAL SOFTWARE. David L. Parnas, A. John van Schouwen and Shu Po Kwan in *Communications of the ACM,* Vol. 33, No. 6, pages 636–648; June 1990.

SOFTWARE SAFETY IN EMBEDDED COMPUTER SYSTEMS. Nancy G. Leveson in *Communications of the ACM,* Vol. 34, No. 2, pages 34–46; February 1991.

FORUM ON RISKS TO THE PUBLIC IN COMPUTERS AND RELATED SYSTEMS. Moderated by Peter Neumann. Available as the usenet newsgroup comp.risks, or by request on the internet from risks-request@csl.sri.com.

RISKS TO THE PUBLIC IN COMPUTERS AND RELATED SYSTEMS. Regular column edited by Peter Neumann in *Communications of the ACM.*

Wire Pirates

Consumers and entrepreneurs crowd onto the information highway, where electronic bandits and other hazards await them

by Paul Wallich

Someday the Internet may become an information superhighway, but right now it is more like a 19th-century railroad that passes through the badlands of the Old West. As waves of new settlers flock to cyberspace in search of free information or commercial opportunity, they make easy marks for sharpers who play a keyboard as deftly as Billy the Kid ever drew a six-gun. Old hands on the electronic frontier lament both the rising crime rate and the waning of long-established norms of open collaboration.

It is difficult even for those who ply it every day to appreciate how much the Internet depends on collegial trust and mutual forbearance. The 30,000 interconnected networks and 2.5 million or more attached computers that make up the system swap gigabytes of information based on nothing more than a digital handshake with a stranger. (Even estimates of the Internet's size, compiled by SRI International, rely on the cooperation of system administrators around the globe.) Most people know, for example, that electronic-mail messages can be read by many people other than their intended recipients, but they are less aware that E-mail and other communications can be almost tracelessly forged—virtually no one receiving a message over the net can be sure it came from the ostensible sender.

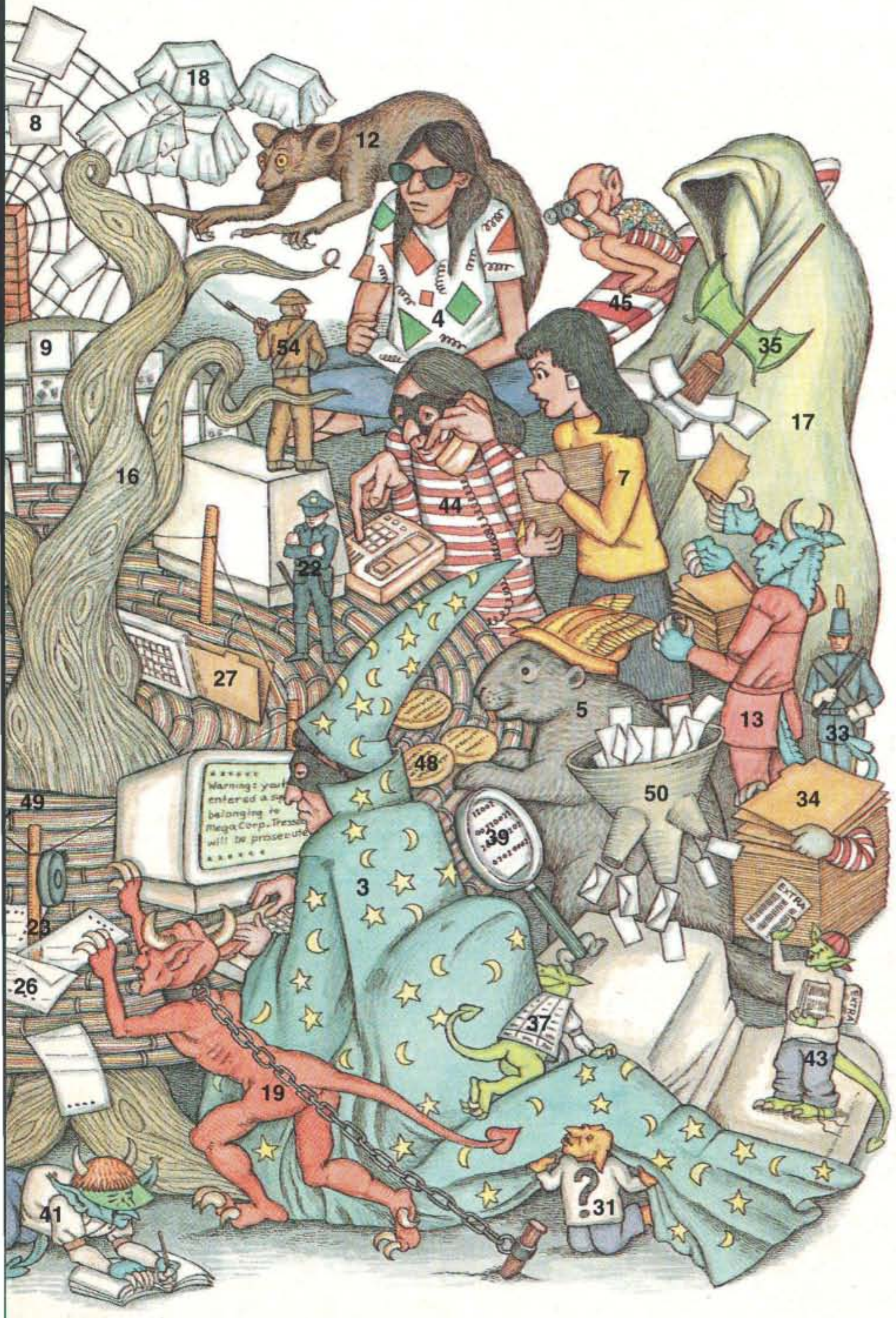
Electronic impersonators can commit slander or solicit criminal acts in someone else's name; they can even masquerade as a trusted colleague to con-

INHABITANTS OF CYBERSPACE may be villains, victims or bystanders (some of them are shown at the right). Self-defense can be difficult in an environment ruled by trust.

Reprinted from the March 1994 issue

vince someone to reveal sensitive personal or business information. Of those few who know enough to worry about electronic forgeries, even fewer understand how an insidiously coded E-mail message can cause some computers to give the sender almost unlimited access to all the recipient's files. And mail-transfer programs are only one of the wide range of ways that an attacker can gain access to a networked computer. "It's like the Wild West," says Donn B. Parker of SRI: "No laws, rapid growth and enterprise—it's shoot first or be killed."

To understand how the Internet, on which so many base their hopes for education, profit and international competitiveness, came to this pass, it can be instructive to look at the security record of other parts of the international communications infrastructure. A computer cracker may become a "phone phreak" to avoid paying for the long-distance habit that computer intrusion sometimes requires, or he may take up phone phreaking as a related hobby (much as a poacher might both hunt and fish). Not only are some of the players the same, so are many of the basic design is-



1. USERS
2. CRACKER
3. WIZARD
4. GURU
5. GOPHER
6. ARCHIE
7. VERONICA
8. WORLD-WIDE WEB
9. MOSAIC
10. FETCH
11. CRON DAEMON
12. FINGER DAEMON
13. FTP (FILE TRANSFER PROTOCOL) DAEMON
14. TELNET DAEMON
15. MAILER-DAEMON
16. ROOT
17. NOBODY
18. ZOMBIE PROGRAMS
19. SATAN
20. CRACK
21. WORM
22. COPS
23. TRIPWIRE
24. TROJAN HORSE
25. PACKET SNIFFER
26. DOT FILES
27. DEVICE FILES
28. SHELLS
29. NETWORK FILE SYSTEM (NFS)
30. DOMAIN NAME SERVER
31. IDENT DAEMON
32. INTERNET-DAEMON
33. LOGIN DAEMON
34. NFS-DAEMON
35. PAGE DAEMON
36. PASSWORD-AUTHORIZATION DAEMON
37. QUOTA DAEMON
38. REMOTE-LOGIN DAEMON
39. STAT-DAEMON
40. SWAPPER
41. SYSLOG DAEMON
42. UPDATE DAEMON
43. USENET NEWS DAEMON
44. FINGER HACKER
45. SHOULDER SURFER
46. CELLULAR-PHONE CLONES
47. REMOTE PROCEDURE CALL
48. MAGIC COOKIES
49. SHELL SCRIPTS
50. FILTER
51. MIME
52. FIRE WALL
53. ROUTER
54. CHALLENGE-RESPONSE SYSTEM
55. KERBEROS

sues. Furthermore, engineers building each new generation of technology appear to make the same mistakes as their predecessors.

The first, biggest error that designers seem to repeat is adoption of the "security through obscurity" strategy. Time and again, attempts to keep a system safe by keeping its vulnerabilities secret have failed. Consider, for example, the running war between AT&T and the phone phreaks. When hostilities began in the 1960s, phreaks could manipulate with relative ease the long-distance network in order to make unpaid telephone calls by playing certain tones into the receiver. (One phreak, John Draper, was known as "Captain Crunch" for his discovery that a modified cereal-box whistle could make the 2,600-hertz tone required to unlock a trunk line.) The precise frequencies were "hidden" in technical manuals and obscure journal articles, but college students and others soon ferreted them out. Phreaks built so-called black, blue and red boxes to produce the required signals, and a small cottage industry flourished until the telephone company adopted methods that were less vulnerable to spoofing through the telephone mouthpiece.

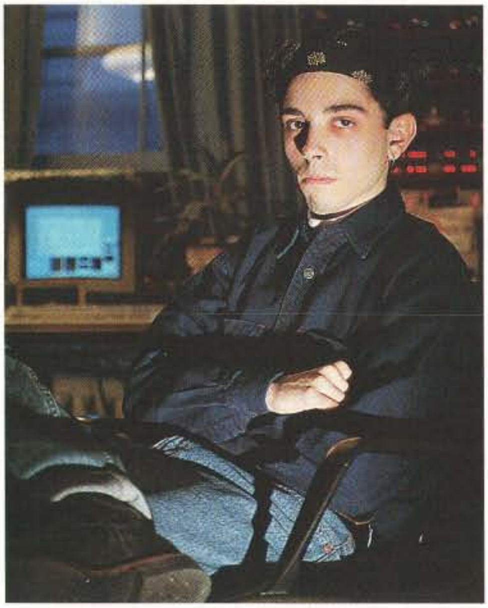
Telephone credit cards underwent an evolutionary process that continues today. When the cards were first introduced, recalls Henry M. Kluepfel of Bell Communications Research, credit-card numbers consisted of a sequence of digits (usually area code, number and billing office code) followed by a "check digit" that depended on the other digits. Operators could easily perform the math to determine whether a particular credit-card number was valid. And phreaks could easily figure out how to generate the proper check digit for any given telephone number. The telephone company had to rely on detecting fraudulent calls as they occurred, tracking phreaks down and prosecuting them, a strategy that "never worked in the long term," according to Kluepfel.

In 1982 AT&T finally put in place a more robust method. The corporation assigned each card four check digits (the "PIN," or personal identification number) that could not be computed easily from the other 10. A nationwide on-line data base made the numbers available to operators so that they could determine whether a card was valid.

Since then, theft of telephone credit-card numbers has become a matter of observation and guile rather than mathematics. "Shoulder surfers" haunt train stations, hotel lobbies, airline terminals and other likely venues. When they see a victim punching in a credit-card number, they transmit it to confederates for widespread use. Kluepfel noted ruefully that his own card was compromised one day in 1993 and used to originate more than 600 international calls in the two minutes before network-security specialists detected and canceled it. "I made a call from a coin phone and shielded the number from the scruffy-looking guy on my left, but I was unaware of the guy in the business suit on my right," he confesses.

Kluepfel cites estimates that stolen calling cards cost long-distance carriers and their customers on the order of half a billion dollars a year. The U.S. Secret Service has placed the total volume of telecommunications fraud at $2.5 billion; industry numbers range from $1 billion to $9 billion.

"It's only insecure if you configure it wrong."
—Mark Abene, "Phiber Optik," hacker



Somebody Else's Problem

Over the course of a generation, AT&T developed monitoring tools to thwart callers trying to evade toll charges. The corporation also used them to foil individuals who dialed into telephone switching systems to manipulate the facilities directly. Such access let phreaks forward other people's telephones to new locations, route calls around the world or even cut off one another's telephone service. After the Bell system breakup in 1984, however, AT&T was no longer the global policeman of the telecommunications world. In particular, tens of thousands of large and small companies that purchased PBXs (so-called private branch exchanges) to automate their internal telephone networks found themselves the targets of "finger hackers," but they had nothing like AT&T's expertise in self-defense.

The simplest way to commit finger hacking, says Kevin Hanley of AT&T, is "dial 1-800 and seven digits." At the other end of many toll-free lines sits a PBX remote-access unit, a subsystem that permits company employees to call their home office and then dial out from there to any number in the world. Most such units require a security code for outgoing calls, Hanley notes, but sometimes "the chairman doesn't want to remember a password." No one knows precisely what such oversight and self-indulgence cost. Industry estimates for PBX fraud range from a few hundred million to more than a billion dollars a year. "It's a bonehead crime," sneers Mark Abene, a hacker now serving a one-year sentence in federal prison for computer intrusion. (During an interview, Abene sketches out the architecture of digital interoffice signaling systems while complaining about what he considers unwarranted slurs on his character in internal telephone-company memos.)

Nevertheless, relentless trial-and-error dialing is highly profitable. Finger hackers can make tens or hundreds of thousands of dollars in calls before being detected. Most such crimes, Hanley notes, are committed by organized groups of criminals, who may even set up storefronts where they sell long-distance calls at cut price. Customers walk in, pay their money and tell an attendant the number they want to call. Less image-conscious thieves may sell calls out of a corner telephone booth.

Almost all these attacks can be circumvented by configuring a system to block calls to locations where a company does no business and by logging incoming calls to detect attempts at intrusion. Yet many PBX owners are unaware of the danger they face, Hanley says. Even those who do know their peril may not have managers on duty at night or during weekends, when most frauds occur. Of the 40 or 50 attendees at the security seminars that Hanley conducts, only "two or three know how toll fraud works," he asserts: "Maybe 30 or 40 percent know what it is, and that it's a bad thing."

Designers of the next big innovation in telecommunications, the cellular telephone, apparently ignored the lessons that their wire-bound predecessors so painfully learned. Leaving aside the fact that anyone with the right radio receiver can listen in on calls, the units are uniquely vulnerable to toll fraud. Every cellular telephone call begins with a broadcast of the telephone's serial number and billing number. Cellular switches check the pair against a data base of working telephones to decide whether a call should go through. Unfortunately, these numbers are also the only information a thief needs to impersonate a legitimate caller.

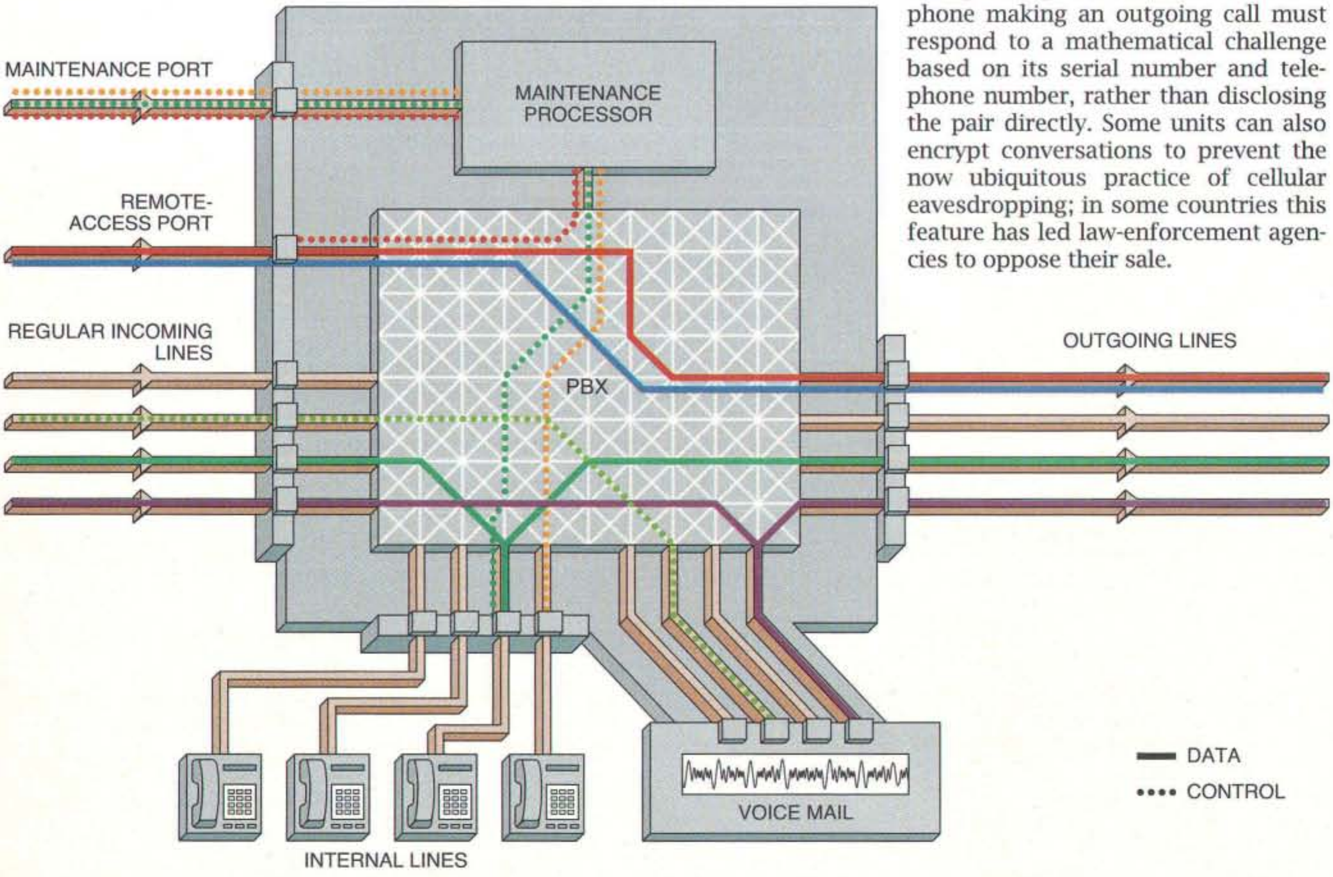
As early as 1984, communications expert Geoffrey S. Goodfellow (who got his start in the field when he broke into a computer at SRI and ended up with a job offer) wrote an article laying out a road map for cellular-phone fraud; only during the past year or two have thieves used the more sophisticated schemes he outlined. The fundamental problem, Goodfellow asserted, was that cellular-phone engineers underestimated both the technical expertise and the persistence of those who might want to subvert their equipment. He called for immediate replacement of current cellular-phone standards with more secure alternatives, but to little avail.

The simplest attack is known as cloning: reprogramming the serial number and telephone number of a pirate unit to that of a telephone currently in use by a legitimate customer. Although standards call for telephones to be built so that it will be impossible to change a serial number without irreparable damage, some early cellular-phone manufacturers supplied the numbers in a memory chip that could be popped out with a screwdriver. Others now place the information in an electrically programmable chip that can be accessed simply by applying the appropriate voltage to the telephone.

More sophisticated frauds adapt the same circuitry that allows a cellular phone to listen for incoming calls to decode the numbers as they are broadcast. A thief can then replay them to make calls that will be billed to others. The companies that carry cellular-phone traffic—and thus are financially responsible for fraudulent charges—have adopted a number of monitoring techniques to detect illicit calls; in addition, some have set their switching equipment to prevent long-distance cellular traffic to areas of the world where fraudsters often call (the countries that topped one carrier's list early in 1993, for example, were the Dominican Republic, Egypt, Pakistan, India, the former Soviet Union, El Salvador, China, Colombia, Mexico and Ghana).

Instead of just monitoring calling patterns, telecommunications engineers have been working on hardware fixes that could block the vast bulk of fraudulent cellular calls. In the fall of 1993 TRW announced a technique for analyzing the analog transmission "signature" of each telephone and storing it along with the serial number and telephone number. If a unit's serial number and telephone number do not match the signature, it must be fraudulent. Details of the characteristics that make up the signature are, of course, supposed to remain secure to prevent evildoers from devising countermeasures.

Such measures are a stopgap. The major cellular carriers are already preparing to replace the current analog cellular-phone system with a digital one. Most proposed digital-cellular standards incorporate protocols in which a telephone making an outgoing call must respond to a mathematical challenge based on its serial number and telephone number, rather than disclosing the pair directly. Some units can also encrypt conversations to prevent the now ubiquitous practice of cellular eavesdropping; in some countries this feature has led law-enforcement agencies to oppose their sale.



TELEPHONE SYSTEMS are vulnerable to a wide range of attacks. "Finger hackers" can dial into a toll-free remote-access port of a PBX and use a security code to make fraudulent calls (*blue*), or they can dupe a voice-mail system into transferring them to an outside line (*purple*). The voice-mail system can be suborned into divulging its contents, transmitting unauthorized messages or destroying stored data. Attackers can also dial into a PBX's maintenance port and disable security on the remote-access port (*red*), forward calls to outside numbers (*green*) or otherwise reconfigure the system.

During the same years that telephone companies were fighting the phone phreaks and cellular-phone architects were designing an estimated billion-dollar annual fraud bill for users and providers of service, computer scientists were laying the foundations of the Internet. Although initial funding came from the Department of Defense's Advanced Research Projects Agency, security was not really a concern, recalls ARPAnet veteran David J. Farber, now at the University of Pennsylvania. In the early days, only researchers had access to the net, and they shared a common set of goals and ethics, points out Eugene H. Spafford of Purdue University.

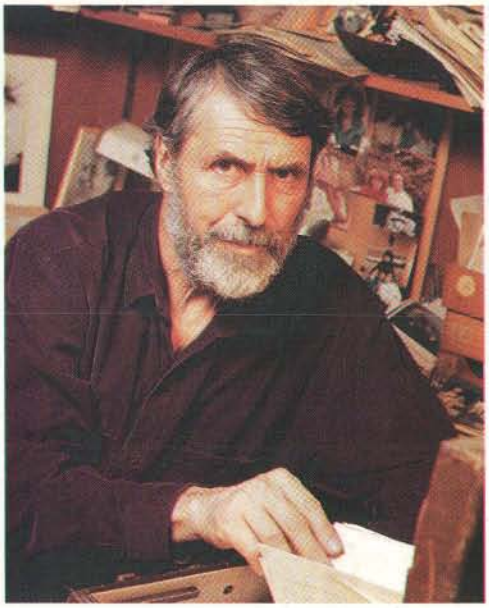
The very nature of Internet transmissions embodies this collegial attitude: data packets are forwarded along network links from one computer to another until they reach their destination. A packet may take a dozen hops or more, and any of the intermediary machines can read its contents. Indeed, many Internet packets start their journeys on a local-area network (LAN), where privacy is even less protected. On a typical LAN, computers broadcast each message to all the other computers attached to the network. Only a gentleman's agreement assures the sender that the recipient and no one else will read the message.

The Cyber-Neighborhood Goes Downhill

A lack of security on the ARPAnet did not bother anyone, because that was part of the package, according to Dorothy E. Denning, a professor of computer science at Georgetown University: "The concerns that are arising now wouldn't have been legitimate in the beginning." As the Internet grew, however, the character of its population began changing, and many of the newcomers had little idea of the complex social contract—and the temperamental software—guiding the use of their marvelous new tool.

By 1988, when a rogue program unleashed by Robert T. Morris, Jr., a Cornell graduate student, brought most Internet traffic to a halt for several days, a clear split had developed between the "knows" and the "know-nots." Willis Ware of Rand, one of the deans of computer security, recalls that "there were two classes of people writing messages. The first understood the jargon, what had happened and how, and the second was saying things like, 'What does that word mean?' or 'I don't have the source code for that program, what do I do?'"

"Crackers are ahead of legitimate administrators; they're way ahead of everybody else."
—Peter G. Neumann, SRI



Since then, the Internet's vulnerabilities have only gotten worse. Peter G. Neumann of SRI, a security researcher who edits the RISKS Forum, an on-line discussion of computer vulnerabilities, characterizes the situation as "unbelievably bad." Anyone who can scrounge up a computer, a modem and $20 a month in connection fees can have a direct link to the Internet and be subject to break-ins—or launch attacks on others. A few years ago the roster counted established names such as "mit.edu," "stanford.edu" or "ibm.com"; today you are as likely to find "mtv.com" or even "pell.chi.il.us," the nom du net of a battered PC-compatible with a hardware-store brass handle screwed into its case for portability.

Moreover, as the Internet becomes a global entity, U.S. laws become mere local ordinances. In European countries such as the Netherlands, for instance, computer intrusion is not necessarily a crime. Spafford complains—in vain, as he freely admits—of computer science professors who assign their students sites on the Internet to break into and files to bring back as proof that they understand the protocols involved.

Ironically, the more thoroughly computerized and networked an organization is, the more risk it may face when making its initial connections to the outside world. The internal network of a high-technology company may look much like the ARPAnet of old—dozens or even hundreds of users, all sharing information freely, making use of data stored on a few central file servers, not even caring which workstation they use to access their files.

As long as such an idyllic little pocket of cyberspace remains isolated, carefree security practices may be defensible. System administrators can safely configure each workstation on their network to allow connections with any other workstation. They can even set up their network file system to export widely used file directories to "world"—allowing everyone to read them—because, after all, the world ends at their corporate boundaries. Indeed, computer companies have made a practice of shipping their wares preconfigured so that each machine automatically shares resources with all its peers.

It does not take much imagination to see what can happen when such a trusting environment opens its digital doors to the Internet. Suddenly, "world" really means the entire globe, and "any computer on the network" means every computer on *any* network. Files meant to be accessible to colleagues down the hall or in another department can now be reached from Finland or Fiji. What was once a private lane is now a highway open to as much traffic as it can bear.

Dan Farmer of Sun Microsystems and Wietse Venema of the Eindhoven University of Technology report that even some of the most respected domains on the Internet contain computers that are effectively wide open to all comers—the equivalent of a car left unattended with the engine running. A recent security alert from CERT (the Computer Emergency Response Team), a clearinghouse for security-related problems based at Carnegie Mellon University, illustrates the point. The alert disclosed that all Sun workstations with built-in microphones had been preset to give their audio input "world-readable" status. Anyone who could gain network access to such a workstation could listen to conversations nearby.

Another alert warned system administrators that the memory buffers that store images displayed on workstation screens might also be preset to world-readable status, as might those that store characters typed on the keyboard. Patient attackers have watched until someone logged in to a privileged account and then simply read the password out of the computer's memory. Once intruders have gained such "root" access, they can masquerade as any legitimate user and read, alter or delete any files. They can also install programs to help them invade other computers and even modify system logs to erase signs of their intrusion.

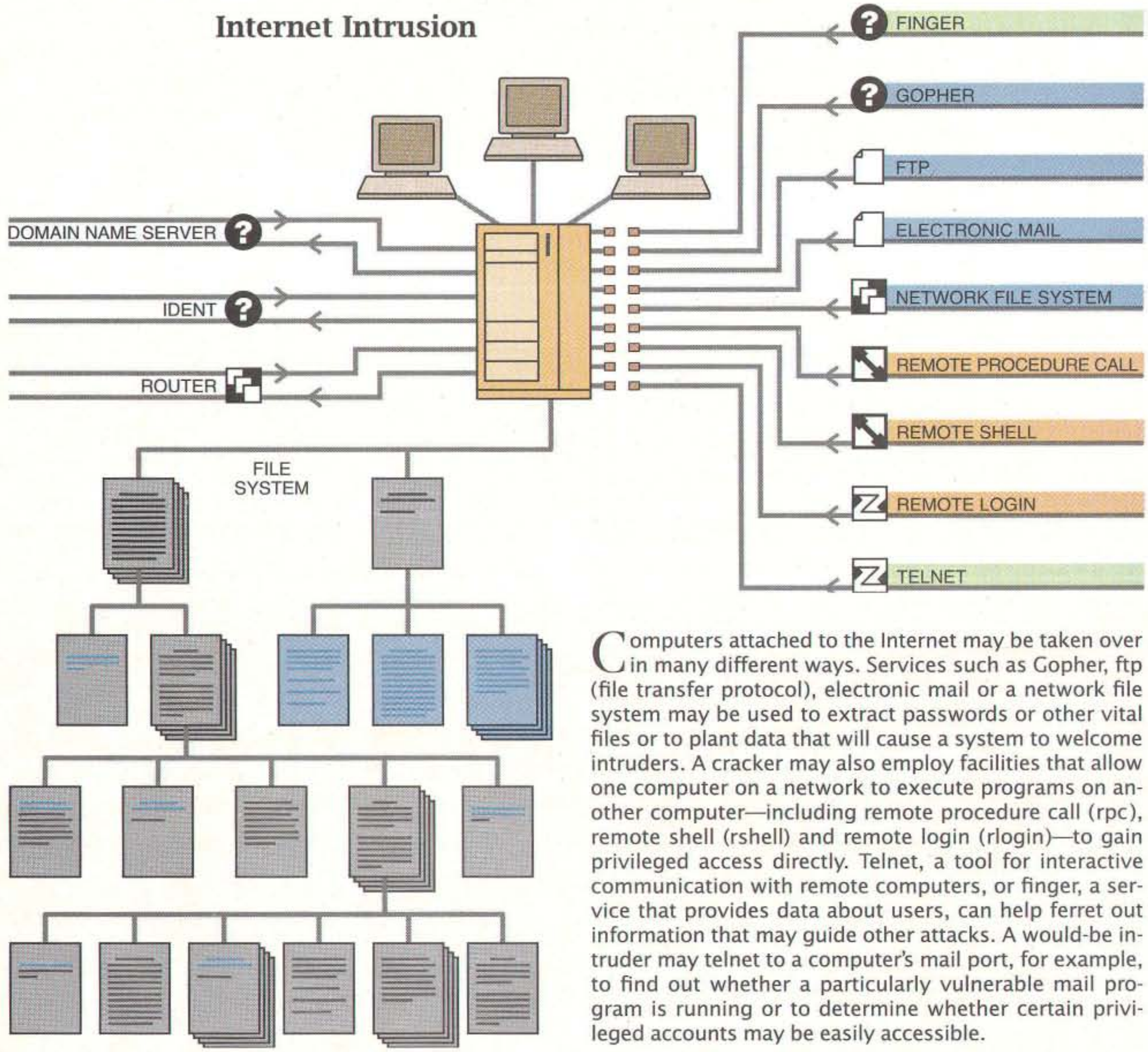
Even if newcomers to the net try to secure their systems, they do not always have an easy time finding the information they need. Computer hardware and software vendors are often loath to talk about security problems, Neumann says. And CERT generally issues net-wide advisories only after manufacturers have developed a definitive fix—weeks or months later, or sometimes never. Most advisories do not explain the security flaw in question; instead they name the software and hardware involved and specify the modifications that should be made to reduce the chances of intrusion.

This policy keeps potentially useful information away from those crackers who are not well connected within the illicit community. But it also keeps many novice system administrators in the dark, Neumann complains: "You find out by having a buddy who's a system administrator somewhere else." Neumann estimates that between half and three quarters of the security holes currently known to hackers have yet to be openly acknowledged. "People don't know the risks," Spafford comments. "They know the benefits because people talk about those."

Many of those benefits come from programs such as Gopher, World-Wide Web or Mosaic, which help people navigate through cyberspace in search of information. A single menu selection or click of a mouse may take a researcher from a computer in Minnesota to another in Melbourne or Zurich. Files containing U.S. census data, pictures of the aft plumbing of the space shuttle and lists of British pubs or of artificial-intelligence software are available free for the finding. Such tools, many of them just a gleam in researchers' eyes a few years ago, are drawing tens of thousands of people to the Internet.

Yet that rapid evolution may have leapfrogged steps that could have contributed to security, Spafford notes. Cybersurfers are relying on "the first or second version" of programs originally written to test ideas rather than to provide industrial-strength services.

The popular Gopher program, according to CERT advisories and other sources, has security flaws that make it possible to access not only public files



Computers attached to the Internet may be taken over in many different ways. Services such as Gopher, ftp (file transfer protocol), electronic mail or a network file system may be used to extract passwords or other vital files or to plant data that will cause a system to welcome intruders. A cracker may also employ facilities that allow one computer on a network to execute programs on another computer—including remote procedure call (rpc), remote shell (rshell) and remote login (rlogin)—to gain privileged access directly. Telnet, a tool for interactive communication with remote computers, or finger, a service that provides data about users, can help ferret out information that may guide other attacks. A would-be intruder may telnet to a computer's mail port, for example, to find out whether a particularly vulnerable mail program is running or to determine whether certain privileged accounts may be easily accessible.

but private ones as well. ("It's only insecure if you configure it wrong," insists Abene, who will return to his job as system administrator for an on-line service in New York City when he is released from prison.) Running Gopher initiates a dialogue between a client program on the user's machine and a Gopher server somewhere on the Internet. The server presents the client with a menu of choices for information, along with a set of "magic cookies"—shorthand specifications for the location of additional information.

If the user wants to delve deeper into a particular subject, the client program sends the server the cookie corresponding to that piece of information. Yet it is relatively easy to modify the cookie so that it specifies the location of information on the server's machine that is supposed to be kept private. An unsuspecting Gopher server will deliver those private files without checking whether the cookie it receives matches one of the items in the menu it presented. Although Gopher servers can readily be confined so that they have access only to public information, by default they have free rein.

By Bits Deceived

Failure to check the propriety of commands is a common oversight—as Othello with Iago, computers on the Internet trust not wisely but too well. E-mail, one of the net's most basic services, sets the tone: an electronic letter consists simply of a text file containing a header specifying the sender, addressee, subject, date and routing information, followed by a blank line and the body of the message. Although mail programs generally fill in the header lines accurately, there is little to prevent a whimsical or malicious person from inserting whatever information they please. A message from "president @whitehouse.gov" could as easily originate from a workstation in Amsterdam as from the Executive Office Building in Washington, D.C. Forging E-mail is "trivial," Farber asserts; what makes such forgeries a problem as the Internet grows is that the incentive for successful forgeries does so, too, as do the dangers of being taken in. Companies and individuals have already begun doing business via E-mail; real money and goods change hands on the basis of electronic promises.

Computer scientists have developed protocols for verifying the source of E-mail messages, but spoofers are also improving their techniques. Correspondence on Usenet discussion groups such as "comp.security.misc" illustrates this coevolution: some security-minded system administrators have advocated the use of "IDENT daemons." If a spoofer connects to a mail server and offers a false identification (the first step in sending a forged message), the mail server can query the IDENT daemon on the spoofer's machine.

"People are attacking individual systems, not the net per se."
—Dorothy E. Denning,
Georgetown University

Others disparage IDENT; they point out that the name returned by the daemon is only as trustworthy as the computer it runs on. Once hackers have gained control of a machine—either by breaking in or because they own it—they can configure the IDENT daemon to respond to queries with whatever name they please.

Some system administrators are meeting the threat of such deceptions by barring connections to their computers from untrustworthy parts of the Internet. Each range of numbered addresses on the Internet corresponds to a particular organization, or domain, and so it is simple to refuse connections from computers in a domain believed to serve as a vehicle for hackers.

Even this step has a countermeasure. Spafford points out that most machines rely on "domain name servers" to translate back and forth between numbered network addresses and domains such as "xerox.com" or "umich.edu." But the name servers are just ordinary computers. They are vulnerable to deception or intrusion, and so the road maps they provide can be rewritten to serve deceitful ends. A cracker can modify the name server's data base so that it tells any computer querying it that the address belonging to, say, "evil.vicious.hackers.org" is instead that of "harvard.edu." A computer that accepts connections from Harvard University will then allow the hackers in as well. Indeed, Spafford laments, it is almost impossible for any program to know for sure where the data packets reaching it over the Internet really come from or where the packets it sends out are going.

Another class of security problems comes not from misplaced trust in domain name servers or IDENT daemons but rather from the same versatility that makes networked computers so useful. Perhaps the best example of this is the "sendmail bug," a disastrous loophole that has reappeared time after time in the history of the net.

The bug arises because most mail programs make it possible to route messages not only to users but also directly to particular files or programs. People forward mail, for instance, to a program called vacation, which sends a reply telling correspondents that the intended recipient is out of town. Many people also route mail through filter programs that can forward it to any of several locations depending on sender, subject matter or content.

But this same mechanism can be subverted to send electronic mail to a program that is designed to execute "shell scripts," which consist of a series of commands to perform system functions, such as extracting information from files or deleting all files older than a certain date. This program will then interpret the body of the message as a script and will execute any commands it contains. Those commands could cause a copy of the receiving computer's password file to be sent to an intruder for analysis, fashion a subtle back door for later entry or simply wreak havoc on the recipient's stored data. Mail sent to certain files can have similar effects.

Some fixes for the latest incarnation of the sendmail bug have been published on the Internet and presumably have been implemented by most system administrators who saw them, but many systems remain vulnerable. Furthermore, other programs that process electronic mail contain analogous holes.

Even more ominous is the fact that E-mail is by no means the sole way to plant uncontrolled data in a victim's computer. Steven M. Bellovin, a researcher at Bell Labs, points out that Gopher and other information-retrieval programs also transfer large, potentially ill-identified files. A hacker would have to go to some trouble to set up a corrupt Gopher server and would even have to stock it with useful information to entice people into making connections to it. "I won't be surprised when it starts happening," Bellovin says.

Walls in Cyberspace

If the Internet, storehouse of wonders, is also a no-computer's-land of invisible perils, how should newcomers to cyberspace protect themselves? Security experts agree that the first layer of defense is educating users and system administrators to avoid the particularly stupid mistakes. People still tape passwords to their keyboards or use no password at all for privileged computer accounts. One graduate student, pressed into service as administrator for a cluster of workstations at the University of Michigan, found that a simple password-guessing program could compromise a quarter of her users' accounts. Five of 80 users had chosen their names as passwords. Some administrators have installed programs that reject passwords based on dictionary words or obvious personal identifiers, but their use is far from widespread.

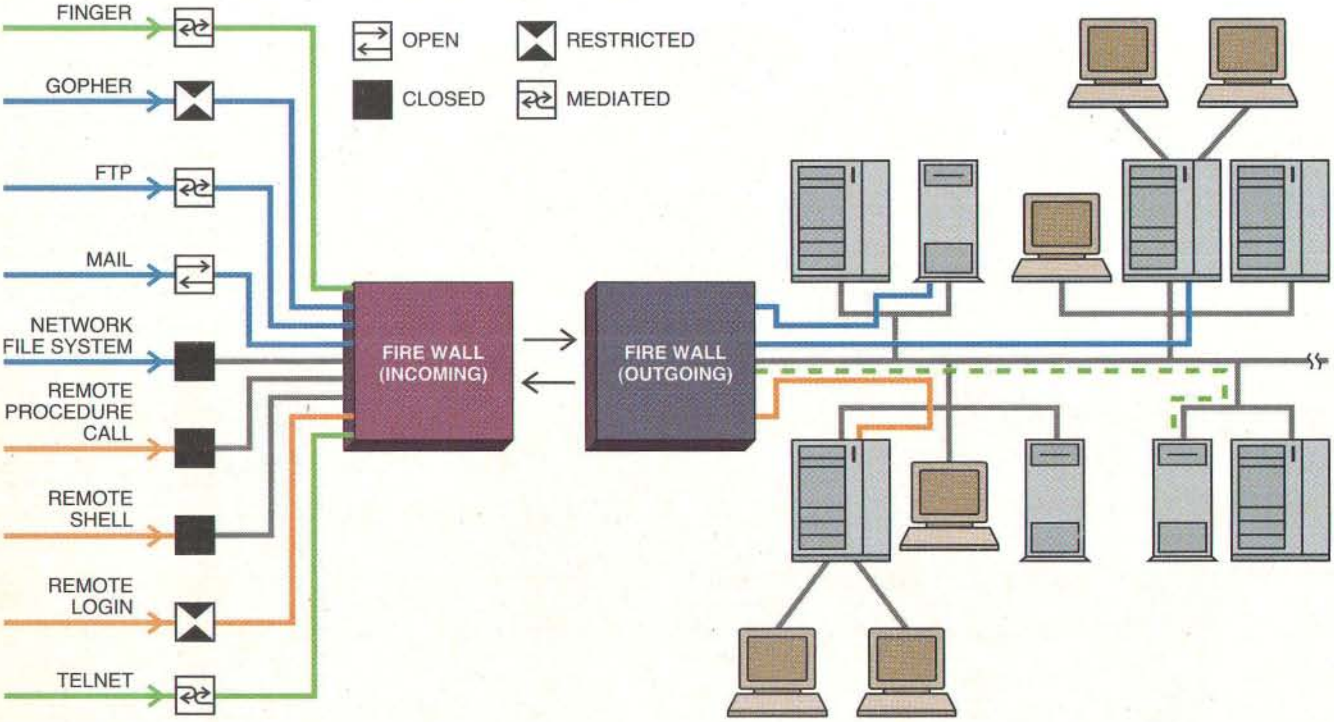
The next level of defense is the so-called fire wall, a computer that protects internal networks from intrusion. Most major companies have long since installed fire walls, and many universities are adopting them as well. Fire walls examine all the packets entering and leaving a domain to limit the kinds of connections that can be made from the Internet at large. They may also restrict the information that can be passed across those connections. "Anyone who would connect a corporate network directly to the Internet should be fired," Farber asserts.

Proposing a fire wall and constructing it are two different matters. Users would like to have access to all possible Internet services. But that desire encounters a harsh reality: "Some things you can't do securely," maintains Marcus J. Ranum of Trusted Information Systems. Ranum, who helped to install the fire wall for "whitehouse.gov," names Gopher and Mosaic as two programs whose trusting nature defies the attempts of a fire wall design to provide safety. In such cases, he argues, security experts must be content to minimize risk rather than eliminate it entirely.

At a bare minimum, a fire wall must pass mail, according to Bellovin (even though mailers may be demonstrably insecure). After that, users want to be able to log in to machines elsewhere on the Internet and to retrieve files from public archive sites or from the directories of colleagues at other institutions.

To perform these functions, AT&T built a fire wall consisting of two dedicated computers: one connected to the Internet and the other connected to the corporation's own network. The external machine examines all incoming traffic and forwards only the "safe" packets to its internal counterpart. In addition, it will accept outgoing traffic only from the internal gateway machine, so an attacker attempting to transfer information illicitly out of AT&T's domain would be unable to do so without subverting the internal gateway. The internal gateway, meanwhile, accepts incoming traffic only from the external one, so that if unauthorized packets do somehow find their way to it, they cannot pass.

Other services are more problematic. Workers would like to be able to log in to their office computers from anywhere on the Internet, for instance. Any intermediate computer relaying traffic over the Internet might have been com-



NETWORK FIRE WALL prevents malicious data packets from wreaking havoc on trusting computers. In some cases, the fire wall may allow access only from trusted locations on the Internet to particular machines inside the fire wall. Or it may allow only demonstrably "safe" information to pass (perhaps permitting users to read the E-mail from remote locations but not to run privileged programs). In other cases, it is impossible to distinguish safe use of a particular network service from unsafe use, and so all requests must be blocked. The fire wall may also provide a substitute for some network services (such as finger, telnet or ftp) that performs most of the same functions but is not as vulnerable to penetration.

promised, however, and could be reading packets (including those containing passwords) as they go by. In fact, in two separate incidents in October 1993, hackers gained access to Panix, a public-access Internet site in New York City, and to BARRNet, an Internet carrier in California, and installed "packet sniffers." These programs watched all the data going by and recorded user names and passwords as people logged in to (at least) hundreds of other computer systems, according to system administrators at Panix.

Such attacks render conventional passwords "obsolete," Ranum asserts. Instead safe connections to machines inside a fire wall require a different kind of authentication mechanism, one that cannot be recorded by a sniffer and then replayed to gain unauthorized access. Two methods are already in limited use: the "one-time" password and "challenge-response."

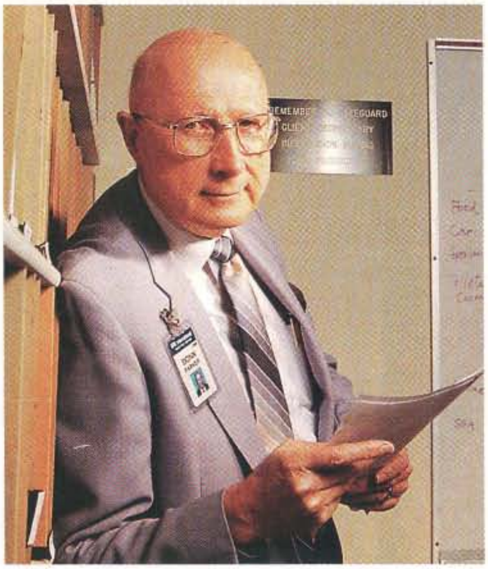
To use one-time passwords, a worker simply carries a list of them. Reuse indicates that an intrusion attempt is in progress. Challenge-response systems have no list of passwords; instead they require an answer to a random query before allowing access. Most often the query consists of a number that must be mathematically transformed by a secret key known only to authorized users. Most people cannot multiply 100-digit numbers in their heads, so commercial challenge-response equipment usually employs a "cryptographic calculator," primed with the key and activated by a shorter sequence that a person can remember.

Encryption Is Key

If passwords should traverse the Internet only in encrypted form, what about other sensitive information? Standardization efforts for "privacy-enhanced" E-mail have been under way for more than five years, but widespread adoption lies well in the future. "The interoperability problem is nasty," Ware says; unless everyone has software that can handle encrypted messages, it is of little use to anyone.

Encryption could provide not only privacy but authentication as well: messages encoded using so-called public-key ciphers can uniquely identify both recipient and sender. But encryption software in general remains at the center of a storm of political and legal controversy. The U.S. government bars easy export of powerful encoding software even though the same codes are freely available overseas.

"At the civil level, constraints. At the criminal level, constraints. At the technological level, constraints."
—Donn B. Parker, SRI



Within the U.S., patent rights to public-key encryption are jealously guarded by RSA Data Security, a private firm that licensed the patents from their inventors. Although software employing public-key algorithms has been widely published, most people outside the U.S. government cannot use it without risking an infringement suit.

To complicate matters even further, the government has proposed a different encryption standard, one whose algorithm is secret and whose keys would be held in escrow by law-enforcement agencies. Although many civil libertarians and computer scientists oppose the measure, some industry figures have come out in favor of it. "You can't have absolute privacy," Parker says. "A democracy just can't operate that way."

The question is not whether cyberspace will be subjected to legislation but rather "how and when law and order will be imposed," Parker says. He predicts that the current state of affairs will get much worse before the government steps in "to assure privacy and to protect the rights people do have."

Others do not have Parker's confidence in government intervention. Ranum foresees an Internet made up mostly of private enclaves behind fire walls that he and his colleagues have built. "There are those who say that fire walls are evil, that they're balkanizing the Internet," he notes, "but brotherly love falls on its face when millions of dollars are involved."

Denning counts herself among the optimists. She lends her support to local security measures, but "I don't lose any sleep over security," she says. Farber, also cautiously optimistic, sees two possible directions for the Internet in the next few years: rapid expansion of existing services, or fundamental reengineering to provide a secure base for the future. He leaves no doubt as to which course he favors. Spafford is like-minded but gloomier. "It's a catch-22," he remarks. "Everyone wants to operate with what exists, but the existing standards are rotten. They're not what you'd want to build on."

Even if computer scientists do redesign the Internet, he points out, putting new standards in place may be impossible because of the enormous investment in old hardware and software. So much of the Internet rests on voluntary cooperation, he observes, that making sweeping changes is almost impossible.

Then again, Ware counters, perhaps piecemeal evolution may be the only possibility. No single organization understands the idea of a national information infrastructure well enough to be put in charge, he contends: "There's no place to go and say, 'Here's the money, work out the problems.' There aren't even three places, and I'm not sure there should be."

In the meantime, the network grows, and people and businesses entrust to it their knowledge, their money and their good names.

FURTHER READING

PRACTICAL UNIX SECURITY. Simson Garfinkel and Gene Spafford. O'Reilly and Associates, 1992.

ZEN AND THE ART OF THE INTERNET. Brendan P. Kehoe. Prentice-Hall, 1993.

IMPROVING THE SECURITY OF YOUR SITE BY BREAKING INTO IT. Dan Farmer and Wietse Venema. Available by ftp from win.tue.nl as /pub/security/admin-guide-to-cracking.Z

FORUM ON RISKS TO THE PUBLIC IN COMPUTERS AND RELATED SYSTEMS. Available via usenet newsgroup comp.risks or by E-mail from risks-request@csl.sri.com. Back issues are available by ftp from crvax.sri.com in the RISKS: directory.

Usenet newsgroups that cover security: comp.security.misc; comp.security. unix; alt.security

Computer Emergency Response Team advisories are available by ftp from cert.org

Security-related information available by ftp from tis.com (Trusted Information Systems) and research.att.com

Surreal Science

Virtual reality finds a place in the classroom

High school physics might have been more fun—and more instructive—if the ability to bounce balls on Jupiter were a laboratory option. Well, now it is, at least in principle. Using his experience in developing virtual-reality environments to train National Aeronautics and Space Administration astronauts, R. Bowen Loftin, a physicist at the University of Houston–Downtown, has developed a virtual-physics laboratory in which this and other improbable exercises are as easy as making a slam-dunk shot on a wastepaper basket.

Students enter virtual reality by donning headgear and a control glove hooked up to computer. A panoramic view of the virtual laboratory is projected through small screens inside the headset. There is a table, a pendulum, some balls as well as a few odd devices that govern the actions in the laboratory. Users see a computer-generated image of their hand, which duplicates the motion of the real thing. Certain gestures of the control glove are reserved for special actions. Pointing the index finger, for example, will send the virtual physicist flying across the room.

In this artificial environment, students can conduct simple experiments in mechanics that are impossible to perform in a real world. Parameters such as friction and drag can be controlled; gravity can even be turned into a negative force. Computer-generated trajectory lines and freeze-time capability enable users to measure the effects of the changes in the variables on swinging pendulums and bouncing balls. Balls, for example, can bounce so that they increase in energy with every skip.

Amusement, however, is not Loftin's primary goal. He and other developers of educational technology are trying to create an interactive environment that challenges students' often mistaken ideas about the physical world. "Kids build misconceptions about the world that are difficult to dislodge," points out Christopher Dede, who directs the center for interactive educational technology at George Mason University.

Loftin notes that many people think constant motion always requires force. This perception—a violation of Newton's first law—comes about because friction almost always plays a role in the motion of objects. "Newtonian mechanics is not easily detected in the world around you," he says.

The immersive quality of virtual reality, Dede says, makes it ideal for such subjects as physics. "We can get a direct intuitive qualitative sense of physical laws. Such simulations are not possible with a mouse running around on a flat screen." Furthermore, the virtual-physics laboratory is not so dangerous as some real physics experiments, as anyone who has gotten a shock from a spark generator or has been bopped by a pendulum knows.

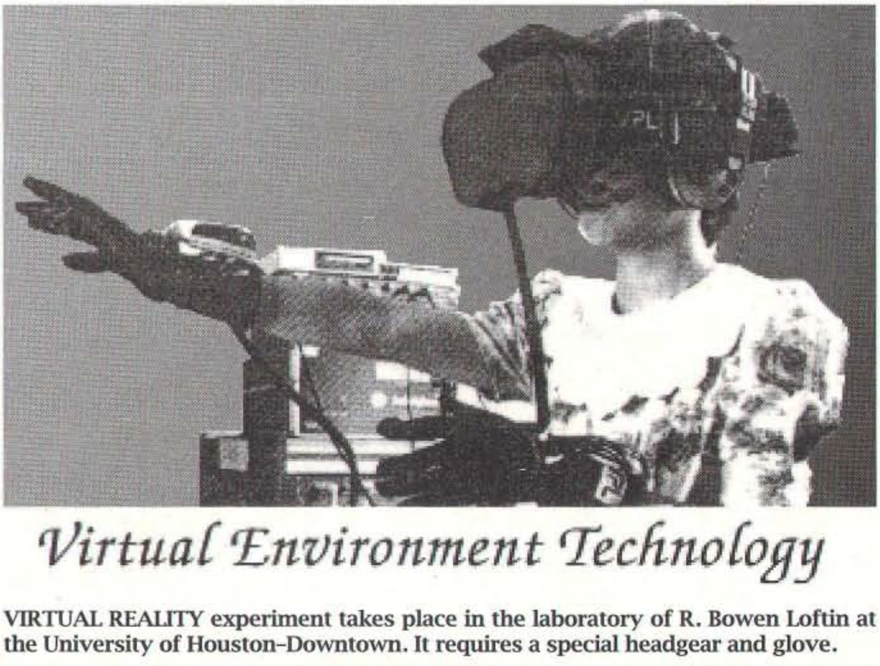
Although most virtual-reality systems retain a certain level of unbelievability, they have proved to be especially useful in situations that would otherwise be difficult to simulate. The near weightlessness of space is one such obvious example. The more conventional techniques all have shortcomings.

In underwater tanks, the viscosity of water affects the training. Freely falling aircraft such as the KC-135 can produce true weightlessness for only about 20 or 30 seconds. Air-bearing floors that levitate astronauts on a cushion of air do not provide the full three-dimensional motion of floating in zero gravity. Virtual reality, on the other hand, can accurately simulate all the conditions of space, so NASA is now using that technology for such tasks as training astronauts to grapple wayward satellites and use robotic arms.

It may take a few years before such computer simulations become commonplace in the classroom. The virtual-physics laboratory is still rather rudimentary. Even by the standards of video games and television cartoons, the images have a stilted quality. Also, the computational speed limits the system's repertoire to a few simple mechanics experiments. But Loftin has been developing a version that will enable users to explore such concepts as angular momentum and energy conservation.

The future could bring virtual laboratories for other domains of science, such as chemistry, biology and medicine. Indeed, Loftin says that medical schools have shown interest in virtual reality for anatomy courses. "It could be like *Fantastic Voyage*," Loftin observes, referring to the movie based on Isaac Asimov's story of scientists shrunk and injected into a human body.

Prices, too, will have to come down before virtual laboratories become widespread. Loftin's headgear and control glove cost several hundred thousand dollars. Researchers are confident, however, that within 10 years, virtual-reality circuit boards should become available for desktop computers. Then, coupled to modest peripheral devices, the fidelity would be sufficient to do some really surreal science. *—Philip Yam*



VIRTUAL REALITY experiment takes place in the laboratory of R. Bowen Loftin at the University of Houston–Downtown. It requires a special headgear and glove.

Electronic Communication

Is it history or just E-Mail?

Computers can act in microseconds, but the law takes a little longer. So the nonprofit National Security Archive in Washington, D.C., discovered when they filed suit in early 1989 to protect from destruction megabytes of data compiled by the Reagan White House staff. The archive contends that the information must be protected under the Federal Records Act. At issue is whether electronic mail constitutes official government records or whether it is simply digital doodling.

The archive, a private group that helps make government information dealing with national security available to scholars, discovered on the eve of George Bush's inauguration that dozens of computer backup tapes—among them ones containing memos from the Iran-Contra fiasco—were about to be erased. Citing federal law that prohibits the wholesale destruction of records, the archive obtained a temporary order preserving the material. The fate of the tapes has been in litigation ever since.

If these memos and E-mail messages were on paper, explains archive general counsel Sheryl Walter, there would be no question that most of them should be preserved. The White House, however, has argued that the documents have no substantive value and fall into the same "nonrecord" category as telephone message slips.

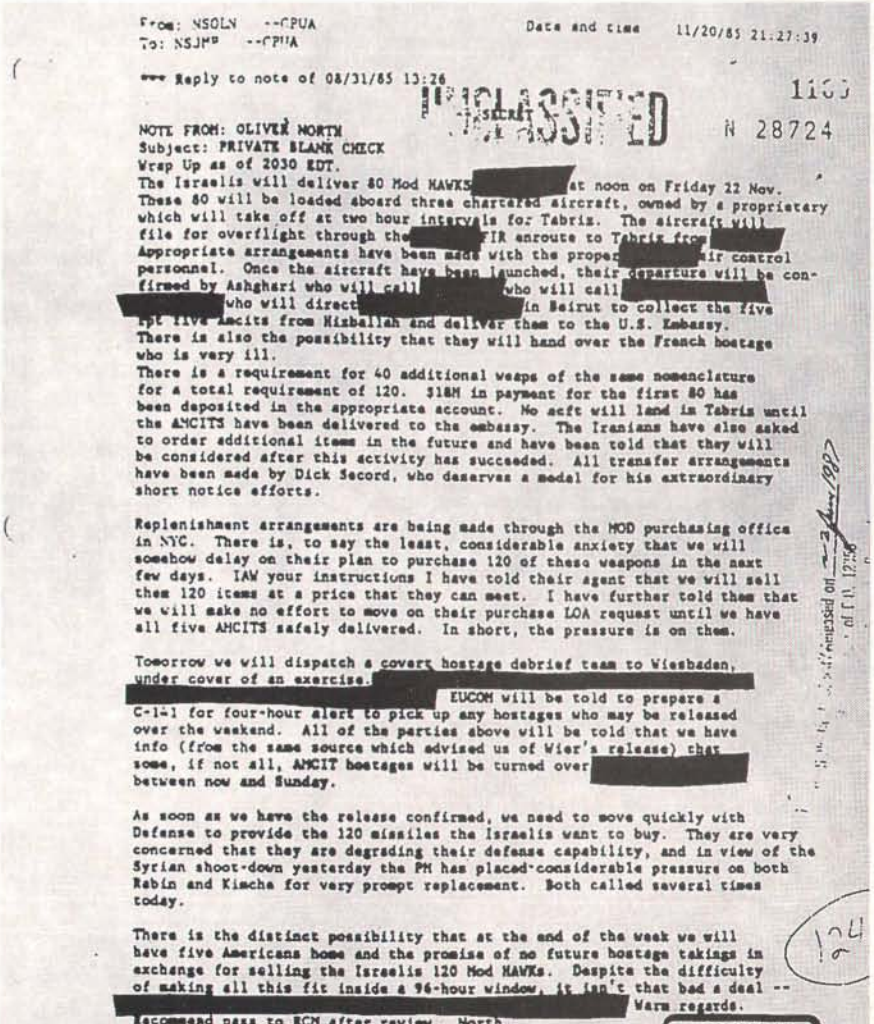
In support of its position, the government cites a policy statement, first committed to paper after the lawsuit was filed, instructing officials that any electronic document attaining "record" status should be printed out on paper and filed for archiving. By implication, anything not printed would be considered ephemeral.

Michael Tankersley, a lawyer at Public Citizen, an advocacy group involved in the case, argues that such a policy leaves individuals with enormous latitude to deny their deliberations to posterity simply by failing to print them out. Furthermore, he points out, the National Security Council allocated only two printers to more than 150 people.

Although high-profile operatives such as Oliver North provide the juiciest potential electronic tidbits, E-mail inhabits a similar limbo at other federal agencies, Tankersley says. At the Department of Energy, for example, the electronic version of officials' personal calendars is considered a record, but E-mail conversations are explicitly unofficial.

Indeed, according to David Bearman of the University of Pittsburgh, the problem goes far beyond the federal government. As business is increasingly transacted by computer, the status of digital documents waxes problematic. Some heavy users of E-mail save everything they send or receive; others discard it all. Managers everywhere, Bearman says, have yet to come to grips with the issue. E-mail and other computerized data are "just out there," he comments. "Managers don't know how they're structured, how they're managed or what happens after they use them."

Federal Judge Charles Richey ruled in early 1993 that E-mail must be preserved. The Clinton administration appealed the decision, but in the summer of 1993 the U.S. Court of Appeals upheld Judge Richey. Nevertheless, Bearman observes, the official or unofficial status of electronic messages will have to be decided by "a larger cultural consensus rather than a narrow legal ruling." *—Paul Wallich*



From: NSOLN --CPUA
To: NSJMP --CPUA

Date and time 11/20/85 21:27:39

*** Reply to note of 08/31/85 13:26

UNCLASSIFIED

SECRET

1163

N 28724

NOTE FROM: OLIVER NORTH
Subject: PRIVATE BLANK CHECK
Wrap Up as of 2030 EDT.
The Israelis will deliver 80 Mod HAWKS [redacted] at noon on Friday 22 Nov. These 80 will be loaded aboard three chartered aircraft, owned by a proprietary which will take off at two hour intervals for Tabriz. The aircraft will file for overflight through the [redacted] FIR enroute to Tabriz from [redacted] Appropriate arrangements have been made with the proper [redacted] ir control personnel. Once the aircraft have been launched, their departure will be confirmed by Ashghari who will call [redacted] who will call [redacted] [redacted] who will direct [redacted] in Beirut to collect the five five Amcits from Hizballah and deliver them to the U.S. Embassy.
There is also the possibility that they will hand over the French hostage who is very ill.
There is a requirement for 40 additional weaps of the same nomenclature for a total requirement of 120. $18M in payment for the first 80 has been deposited in the appropriate account. No acft will land in Tabriz until the AMCITS have been delivered to the embassy. The Iranians have also asked to order additional items in the future and have been told that they will be considered after this activity has succeeded. All transfer arrangements have been made by Dick Secord, who deserves a medal for his extraordinary short notice efforts.

Replenishment arrangements are being made through the MOD purchasing office in NYC. There is, to say the least, considerable anxiety that we will somehow delay on their plan to purchase 120 of these weapons in the next few days. IAW your instructions I have told their agent that we will sell them 120 items at a price that they can meet. I have further told them that we will make no effort to move on their purchase LOA request until we have all five AMCITS safely delivered. In short, the pressure is on them.

Tomorrow we will dispatch a covert hostage debrief team to Wiesbaden, under cover of an exercise. [redacted] EUCOM will be told to prepare a C-141 for four-hour alert to pick up any hostages who may be released over the weekend. All of the parties above will be told that we have info (from the same source which advised us of Wier's release) that some, if not all, AMCIT hostages will be turned over [redacted] between now and Sunday.

As soon as we have the release confirmed, we need to move quickly with Defense to provide the 120 missiles the Israelis want to buy. They are very concerned that they are degrading their defense capability, and in view of the Syrian shoot-down yesterday the PM has placed considerable pressure on both Rabin and Kimcha for very prompt replacement. Both called several times today.

There is the distinct possibility that at the end of the week we will have five Americans home and the promise of no future hostage takings in exchange for selling the Israelis 120 Mod HAWKs. Despite the difficulty of making all this fit inside a 96-hour window, it isn't that bad a deal -- [redacted] Warm regards.
Recommend pass to RCM after review. North.

UNCLASSIFIED

Declassified on 3 June 1987

124

EXHIBIT
CNW-6

RESCUED COMMUNICATION from Oliver North to John Poindexter was retrieved from White House backup files. It was originally sent as E-mail of the kind the Bush administration contended did not have to be preserved.

150 YEARS OF INNOVATION SCIENTIFIC AMERICAN

Special Issue. Order Now.

Sponsored by Apple

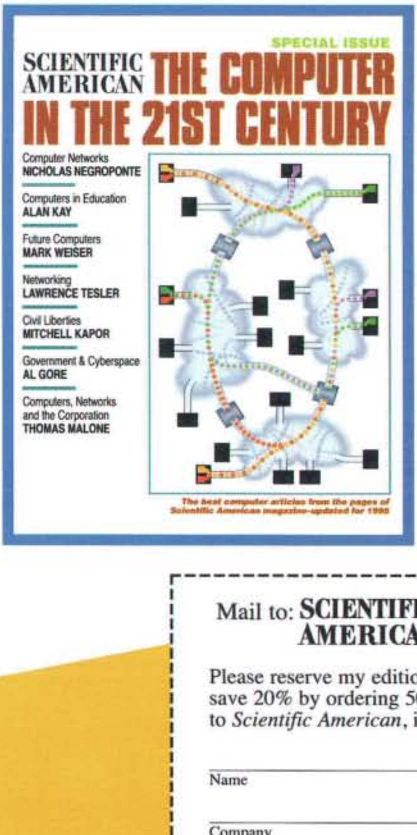
"The Computer in the 21st Century"

If you're a traveler along the world's information superhighway – or anticipate venturing down that road – be sure to use this trail-blazing guide from *Scientific American*: "The Computer in the 21st Century."

This unique anthology features 32 acclaimed *Scientific American* articles by the world's leading experts, including:

- Computers, Networks and Work by *Lee Sproull, Boston University* and *Sara Kiesler, Carnegie Mellon University*
- Networked Computing in the 1990s by *Lawrence G. Tesler, Apple Computer Inc.*
- The Computer for the 21st Century by *Mark Weiser, Xerox Palo Alto Research Center*
- Lightwave Communications by *Emmanuel Desurvire, Columbia University*

"The Computer in the 21st Century" features over 125 pages of information on lightwave communications... microlasers... cyberspace... virtual reality... quantum cryptography... digital technology... and so much more. All beautifully illustrated with the colorful art and graphics for which *Scientific American* is renowned.



SPECIAL ISSUE

SCIENTIFIC AMERICAN THE COMPUTER IN THE 21ST CENTURY

Computer Networks
NICHOLAS NEGROPONTE

Computers in Education
ALAN KAY

Future Computers
MARK WEISER

Networking
LAWRENCE TESLER

Civil Liberties
MITCHELL KAPOR

Government & Cyberspace
AL GORE

Computers, Networks and the Corporation
THOMAS MALONE

The best computer articles from the pages of Scientific American magazine–updated for 1995

Don't let the future pass you by – request your copy of "The Computer in the 21st Century" by returning the coupon today with your remittance. Consider ordering several issues – they make great gifts for friends and colleagues.

SAVE 20% ON BULK ORDERS

Reserve Your Copies Now!

Mail to: **SCIENTIFIC AMERICAN** Dept. CA, P.O. Box 11314
Des Moines, IA 50340-1314

Please reserve my edition(s) of "The Computer in the 21st Century." I'll save 20% by ordering 50 or more copies. My check or money order, payable to *Scientific American*, is enclosed.

Name (PLEASE PRINT)

Company

Address Apt./Suite

City State Zip

Please send _______ copy(ies) at $3.95 each	$________
Less 20% on orders of 50 or more	$________
SUBTOTAL	$________
Plus $1 per copy postage & handling (1-9 copies)	$________
TOTAL AMOUNT ENCLOSED	$________

Please remit in U.S. funds drawn on a U.S. bank. Canadian residents add 7% G.S.T. #R127387652. Quebec residents add Q.S.T. Photocopies of this coupon are acceptable.

PBCAW5B

The future is better than you expected.

Ten short years ago, when Apple introduced the first Macintosh® personal computer, the scientific and technical communities were among the first to embrace its virtues.

To this day, you remain one of our most vital markets. One that we turn to often for inspiration when developing the very innovations outlined in this publication.

And it is because of you that the future has never looked better.

In the preceding pages, we've tried to give you a glimpse of where technology will be moving in the not so distant future.

We've told you how Power Macintosh™ and RISC microprocessor architecture are changing the face of the computer industry, delivering a totally new performance standard that surpasses all previous personal computing standards.

We've told you how software written to take advantage of their unique capabilities will blow the doors off the same programs written for other personal computers.

We've told you how we will continue to break down the barriers in compatibility between operating systems and applications.

And we've told you how we're creating alliances with some of our industry's largest and most innovative companies to build a platform that will open the way to the computers of tomorrow, while enhancing the capabilities of whatever's sitting on your desk right now.

At Apple, we continue to practice a philosophy that most computer companies still haven't managed to grasp: that technology is only useful if it helps people bridge the gap between where they are and where they want to be.

Which may be why Macintosh has consistently been rated significantly higher for user productivity than all other kinds of computers.*

To that end, we will always be committed to building computers that anticipate the way people work, instead of forcing people to adapt to the way computers work.

To providing you with tools that make it easier to do everything you have to do. Yet give you the freedom to do things you never thought you could do before.

To giving you the most useful kind of power a computer can give.

The power to be your best.®

Apple

---

*1994 Arthur D. Little survey.

©1995 Apple Computer, Inc. All rights reserved. Apple, the Apple logo, Macintosh and "The power to be your best" are registered trademarks of Apple Computer, Inc. Power Macintosh is a trademark of Apple Computer, Inc.